第九章　回归分析

§9.1　回归分析模型

在我们的工作学习中，经常需要研究事物之间的关系。这些研究大多数可以量化，

在量化描述事物时有两类关系：

- 确定关系
- 相关关系

如：知道正方形的边长 L 便可知它的面积 $S = L^2$，这种可用确定的函数关系表示的关系称为确定关系；

又如：产品的价格与需求量之间存在某种联系，但无法用确定的函数来描述，这种关系称为相关关系

人的体重与身高的关系，某农作物的亩产量与其播种量，施肥量之间的关系，经济和人口的关系，房价和收入的关系都是相关关系

上述各量之间，可能受别的量影响，甚至在提取出主要变量后，还可能涉及一些难以控制的变量。

问题：如何处理这些难以控制的变量？

方法：将难以控制的变量视为随机变量，追求主要变量间近似的函数关系。

回归分析的目的之一：

寻求具有相关关系的变量之间近似的函数关系

记考察的目标为因变量Y，而影响它的因素为自变量X_1，X_2，…X_k。

问题：怎样描述因变量Y与自变量X_1，X_2，…X_k之间的函数关系呢？

实际中，常把因变量 Y 看作随机变量，自变量X_1，X_2，…X_k看作可控非随机变量。(注：仍然是变量！)

在多元回归模型中最简单的为一元回归模型：

$$y = \mu(x) + \varepsilon, \quad E(\varepsilon) = 0, \ D(\varepsilon) = \sigma^2$$

相应的一元回归方程为： $y = \mu(x)$

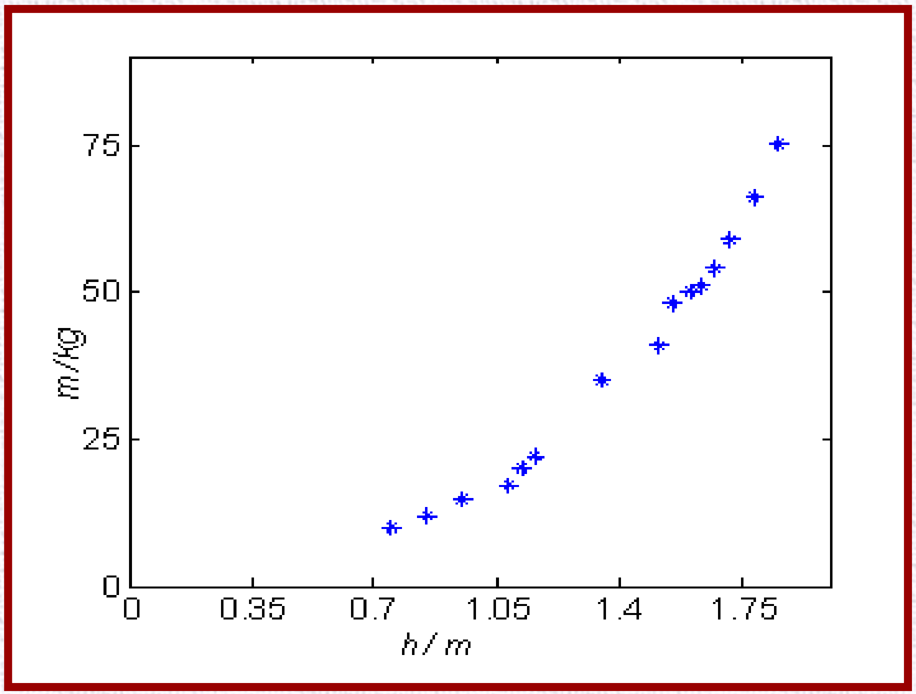
如何确定回归函数$\mu(x)$？

参数回归：由经验或专业知识先确定$\mu(x)$的数学形式，再通过变量观测值(数据、样本)估计未知参数，从而确定$\mu(x)$.

建立模型 → 确定$\mu(x)$的类型 → 估计参数

例 身高体重关系

现有15对某地区人的身高h和体重数据m，希望用简洁的函数关系式描述该地区人的身高体重的对应关系.



呈现幂函数的增长趋势，可设

$$m = \hat{\mu}(h) = b h^{a}$$

其中a, b是待定参数.

注: $\ln h$和$\ln m$和呈线性关系

例 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,试分析施肥量与土豆产量间关系.

施肥量	产量
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

N

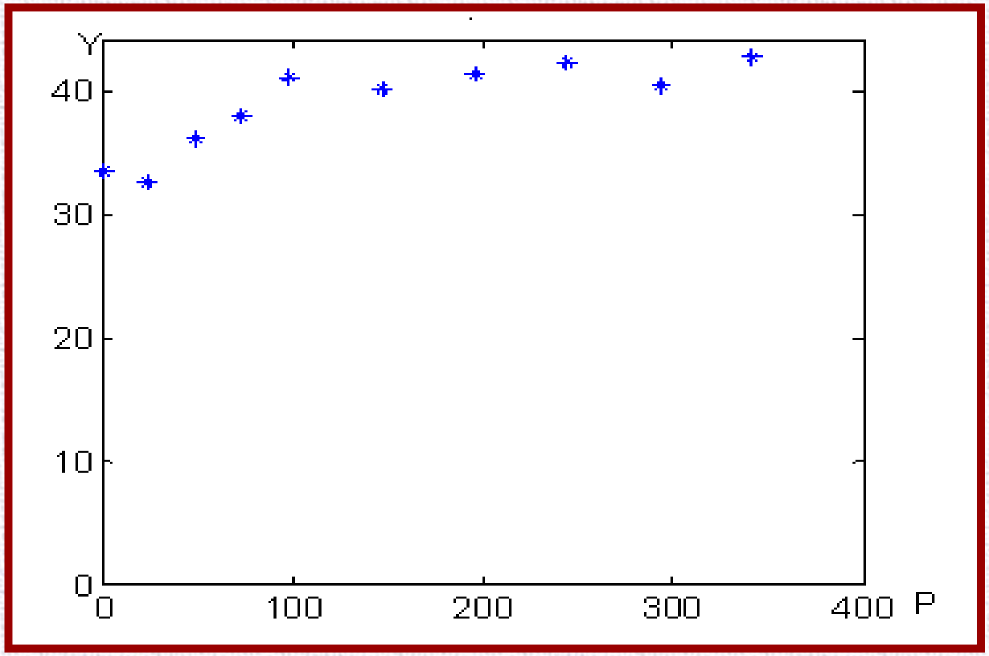
施肥量	产量
0	34.46
24	32.47
49	36.06
73	37.96
98	41.04
147	40.09
196	41.26
245	42.17
294	40.36
342	42.73

P

施肥量	产量
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

K

土豆产量——磷肥量数据散布图



可选 $y=\hat{\mu}(x)=\dfrac{1}{a+be^{-x}},\ x\geq 0.$ a, b是待定参数.

注1： 上述设定仅为初步感性认识，还需进行检验.

注2： 多数非线性关系等价于线性关系

§9.2 一元线性回归分析

一元正态线性回归模型：

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

其中　a——回归常数(又称截距)

b——回归系数(又称斜率)　ε——随机扰动项

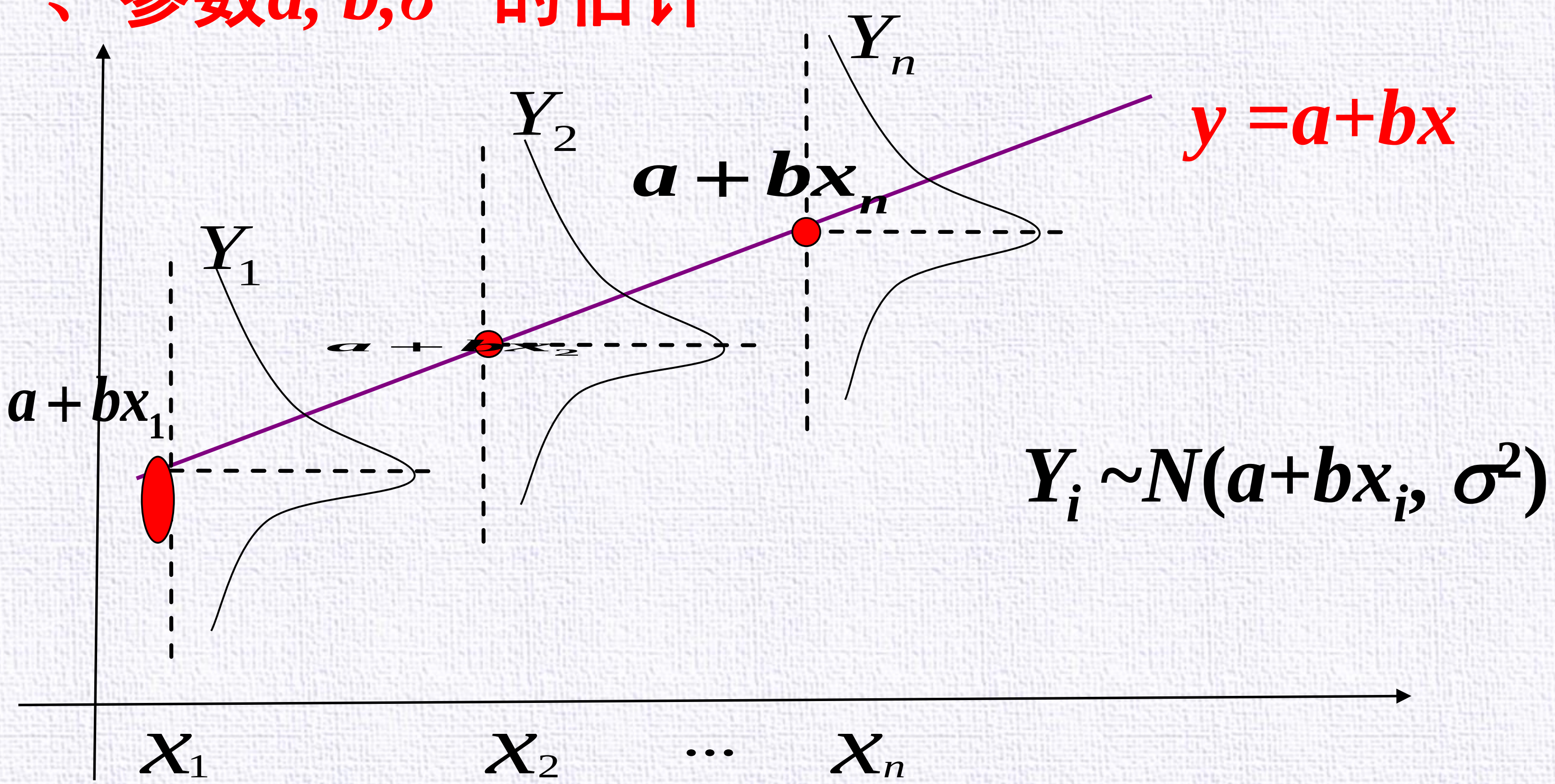
取定自变量X的一组值x_1，x_2，… x_n，进行 n 次独立试验，记结果为Y_1，Y_2，… Y_n。则

$$Y_i = a + bx_i + \varepsilon_i,\quad i=1,\dots,n$$

由各次试验相互独立知 ε_1，…，ε_n相互独立且$\varepsilon_i\sim N(0, \sigma^2)$。从而，若$x_i$已知(注：$x_i$为变量)

$Y_i \sim N(a+bx_i, \sigma^2)$，且相互独立，其形状如图：

一、参数a, b, σ^2 的估计



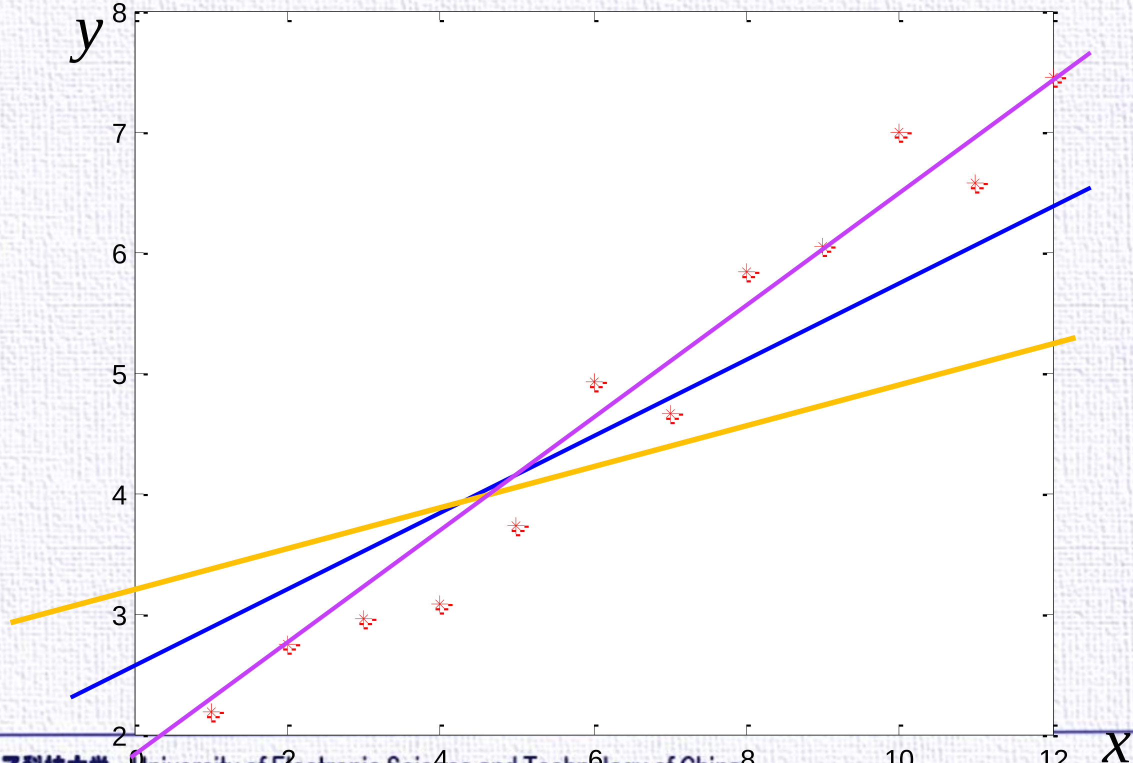
记$\hat{y}_i = a + bx_i$为Y_i的估计值，则 $Y_i = a + bx_i + \varepsilon_i = \hat{y}_i + \varepsilon_i$

这可写成：$\varepsilon_i = Y_i - \hat{y}_i = Y_i - (a + bx_i) \sim N(0, \sigma^2)$

这表明ε_i是Y的实际值与估计值之差，即拟合误差。

若在具体试验中得到自变量X与因变量Y的一组观测值：

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	2.2	2.7	2.9	3.1	3.7	4.9	4.6	5.8	6	7	6.6	7.4

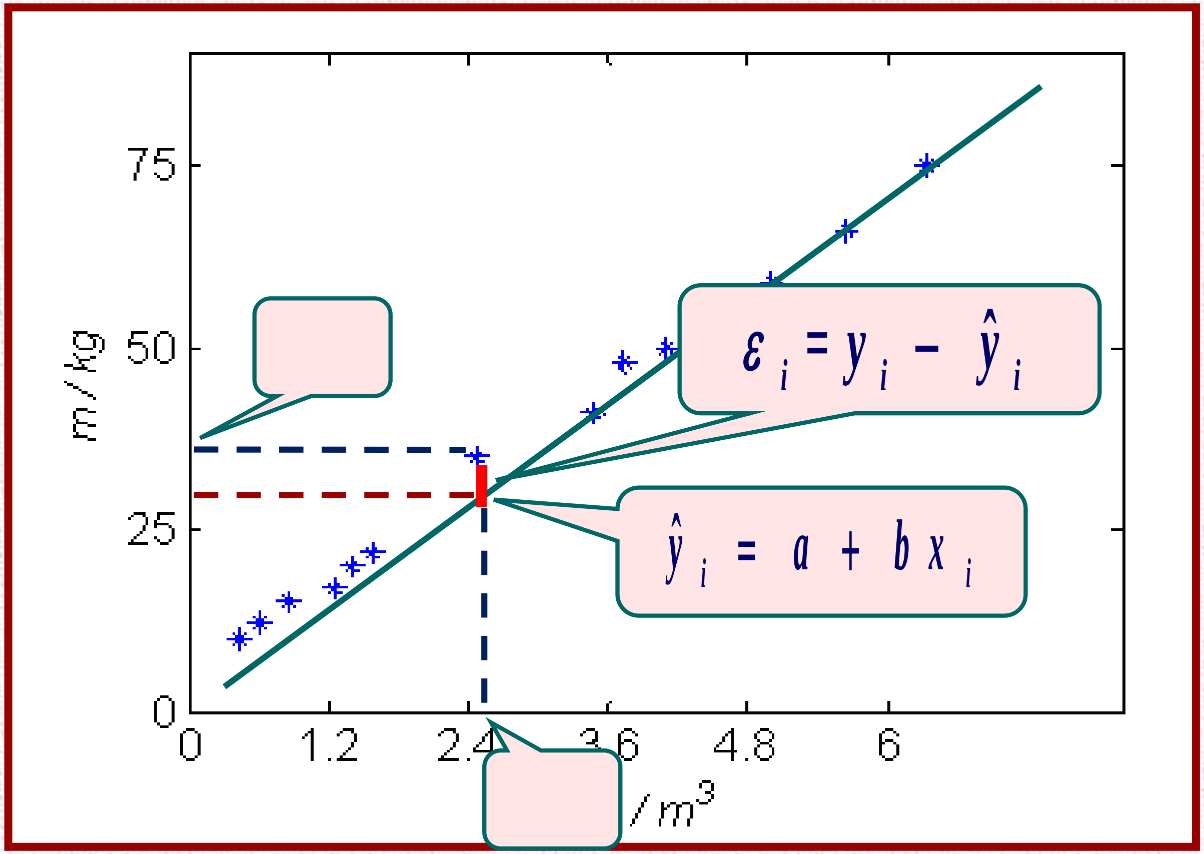


如何确定两个变量之间近似的线性函数关系？

即确定模型

$Y=a+bx+\varepsilon$,

中的a和b？



y_i 的估计值

$\hat{y}_i = a + bx_i$

称为**回归值**.

对所有的i，应使偏差 $y_i - \hat{y}_i$ 都尽可能小，即要使Y的估计值尽可能接近真实值，有三种思路：

(1) 误差总和 $\sum_{i=1}^{n}\varepsilon_i=\sum_{i=1}^{n}(y_i-\hat{y}_i)$ 最小

缺点:可能正负误差抵消

(2) 误差绝对值之和 $\sum_{i=1}^{n}|\varepsilon_i|=\sum_{i=1}^{n}|y_i-\hat{y}_i|$ 最小

缺点：数学处理困难

(3) 误差平方和 $\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 最小 √

结论:应选a,b的估计值，使残差(误差)平方和

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-(a+bx_{1i})]^2$$

达到最小.

一元线性回归分析

因变量Y与关于自变量X的线性回归模型为

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

估计问题变为：确定a和b的估计值，使得

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-(a+bx_{1i})]^2$$

达到最小.

注：y_i与x_{1i}为Y和X_1的数据

建立代数模型

出发点：Q的形式与向量间的欧式距离相似，因此先定义相应向量。

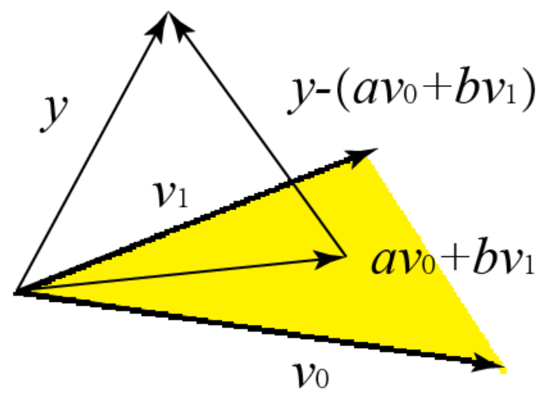
$$y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, A=\begin{pmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix}=(v_0 \quad v_1), r=\begin{pmatrix} a \\ b \end{pmatrix}$$

根据上述定义，最小二乘法可表示为：

寻找$\hat{\boldsymbol{r}}=(\hat{\boldsymbol{a}},\hat{\boldsymbol{b}})$使得

$$\|\boldsymbol{y}-\boldsymbol{A}\hat{\boldsymbol{r}}\|=\min_{\boldsymbol{a},\boldsymbol{b}}\|\boldsymbol{y}-\boldsymbol{a}\boldsymbol{v}_0-\boldsymbol{b}\boldsymbol{v}_1\|$$

代数模型的几何背景



问题回看：寻找$\hat{r} = (\hat{a}, \hat{b})$使得

$$\|y - A\hat{r}\| = \min_{a,b} \|y - av_0 - bv_1\|$$

答案：当$y - av_0 - bv_1$垂直于v_0和v_1张成的平面时，距离最近

数学推导

$$\|y - A\hat{r}\| = \min_{a,b} \|y - av_0 - bv_1\|$$

$$\Leftrightarrow \quad y - av_0 - bv_1 \text{垂直于} v_0, v_1$$

$$\Leftrightarrow \quad v_i^T(y - av_0 - bv_1) = 0$$

$$\Leftrightarrow \quad A^T(y - Ar) = 0$$

$$\Leftrightarrow \quad A^T A r = A^T y$$

若矩阵A列满秩，则A^TA满秩。此时最小二乘解存在唯一，即最小二乘解为

$$\hat{r} = (A^TA)^{-1}A^Ty$$

对一元线性回归问题

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

矩阵A为

$$A = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix}$$

微积分法求最小二乘估计：

通过求Q的最小值来估计a、b

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

要使Q达到最小，应使Q对应于a、b的一阶偏导为0

$$\begin{cases}\dfrac{\partial Q}{\partial a}=-2\sum_{i=1}^{n}(y_i-a-bx_i)=0\\[2ex]\dfrac{\partial Q}{\partial b}=-2\sum_{i=1}^{n}(y_i-a-bx_i)x_i=0\end{cases}$$

上述方法还需要讨论二阶导，并且在多元时，对于包含向量的方程组略显麻烦

可化为: $\begin{cases} na + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i & (1) \\ a\sum\limits_{i=1}^{n} x_i + b\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i & (2) \end{cases}$

$(2)\times n - (1)\times \sum\limits_{i=1}^{n} x_i$得:

$$b[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2] = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

故 $$\hat{b} = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - (\sum\limits_{i=1}^{n} x_i)^2} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

由(1)得a的最小二乘估计值: $\hat{a} = \frac{\sum\limits_{i=1}^{n} y_i}{n} - \hat{b}\frac{\sum\limits_{i=1}^{n} x_i}{n} = \overline{y} - \hat{b}\overline{x}$

估计量的统计性质

问题：估计量从何而来？前面的计算似乎没有涉及到随机。

回顾：实际中，常把因变量 Y 看作随机变量，自变量 X_1，X_2，… X_k 看作可控非随机变量。

一元正态线性回归模型：

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

问题重述：有了一组自变量的取值，假设Y服从上述分布。拿到Y的取值后，估计参数。

最小二乘估计值：

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

最小二乘估计量：

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

注意： $Y_i \sim N(a + bx_i, \sigma^2)$

$$E(\hat{b}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(Y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = b$$

同理： $E(\hat{a}) = a$

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

σ^2的矩估计为 $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$

由于这是有偏估计，将其修正为无偏估计得

P189证明

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \bar{y} + \hat{b}\bar{x} - \hat{b}x_i)^2 \\
&= \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{b}^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&\triangleq \frac{1}{n-2}(l_{yy} - \hat{b}^2 l_{xx})
\end{aligned}$$

$$l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$$

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

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

$$\hat{b}=\frac{l_{xy}}{l_{xx}},\quad \hat{a}=\bar{y}-\hat{b}\bar{x},$$

$$\hat{\sigma}^2=\frac{1}{n-2}(l_{yy}-\hat{b}^2 l_{xx})$$

由试验数据估计得到

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

Y 对 X 的经验回归方程

可证明 $\hat{a}, \hat{b}, \hat{\sigma}^2$ 分别是 a, b, σ^2 的无偏估计，且当 $\varepsilon\sim N(0,\sigma^2)$ 时，也都服从正态分布（参见教材P189）

二、一元线性回归的显著性检验（相关系数法）

问题： 变量Y与X间是否存在线性相关关系？

相关系数法： 基于试验数据检验变量间线性相关关系是否显著的一种方法。

相关系数 $$\rho_{XY}=\frac{E\{[X-E(X)][Y-E(Y)]\}}{\sqrt{D(X)}\sqrt{D(Y)}}$$

是表征随机变量Y与X的**线性相关程度**的数字特征。

样本相关系数：

$$\hat{\rho}_{XY}=R=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

直观： R和相关系数很接近，如果$|R|$较大，就存在线性关系

结论：

1）$|R|$越接近于1，X与Y间的线性相关关系越显著；

2）$|R|$越靠近于0，X与Y间的线性相关关系越不显著。

假设检验：原假设H_0：$\rho_{XY}=0$

给定显著性水平α

当$|R|\le R_\alpha(n-2)$，接受原假设

认为X与Y之间的线性相关关系不显著。

当$|R|>R_\alpha(n-2)$，拒绝原假设

认为X与Y之间的线性相关关系显著。

注：只关心$|R|$是否够大，不需要去考虑左右两侧概率为$\alpha/2$的区间。

注：$R_\alpha(n-2)$的定义查书P192，取值查最后一页

例： 流经某地区的降雨量 X 和该地河流的径流量 Y的观察值如下表，求 Y 关于 X 的线性回归方程。

降雨量 x_i:	110	184	145	122	165	143	78	129	62	130	168
径流量 y_i:	25	81	36	33	70	54	20	44	1.4	41	75

降雨量总和 1436 径流量总和 480.4

解： $n=11$, $\bar{x}=1436/11\approx 130.5$, $\bar{y}=480.4/11\approx 43.7$

$$l_{xx}=\sum_{i=1}^{11}(x_i-\bar{x})^2=13768.7$$

$$l_{xy}=\sum_{i=1}^{11}x_iy_i-n\bar{x}\,\bar{y}=71424.8-62731.35=8693.45$$

$$\begin{cases} \hat{b} = \dfrac{l_{xy}}{l_{xx}} = \dfrac{8693.45}{13768.7} = 0.63 \\ \hat{a} = \overline{y} - \hat{b}\overline{x} = 43.7 - 0.63 \times 130.5 = -38.5 \end{cases}$$

所求经验回归方程为

$$\hat{y} = \hat{a} + \hat{b}x = 0.63x - 38.5$$

$$l_{yy} = \sum_{i=1}^{n}(y_i - 43.7)^2 = 6050.59$$

随机误差的方差σ^2的估计值为

$$\hat{\sigma}^2 = \frac{1}{n-2}(l_{yy} - \hat{b}^2 l_{xx})$$

$$=（6050.59-0.63^2 \times 13768.7）/ \ 9$$

（续前例）用相关系数显著性检验法,检验降雨量 X 和径流量 Y 的线性相关关系是否显著($\alpha=0.01$)。

解：X 与 Y 的样本相关系数为

$$R=\frac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}}$$

$$=\frac{8693.45}{\sqrt{13768.7}\sqrt{6050.58}}=0.952$$

查表得

$$R_\alpha(n\text{-}2)=R_{0.01}(9)=0.735<0.952=R$$

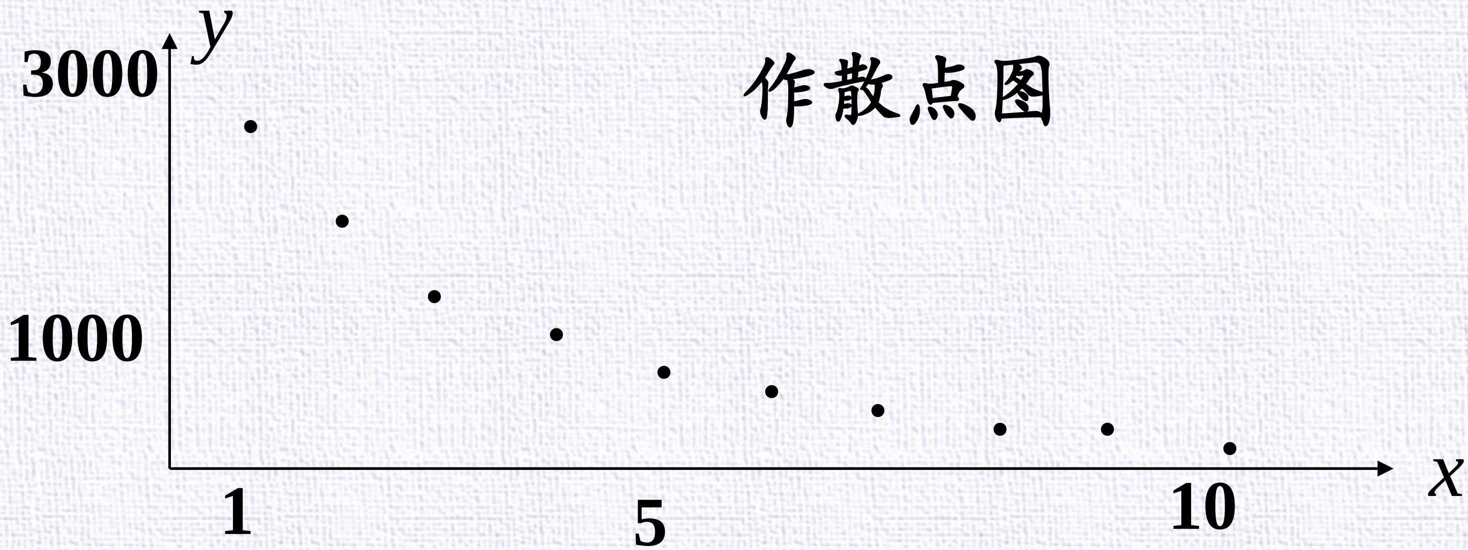
可认为 X 与 Y 的线性相关关系显著。

例：下表是1957年美国旧轿车的调查数据表

使用年数x_i与平均价格y_i（i=1, ...10）

x_i	1	2	3	4	5	6	7	8	9	10
y_i	2651	1943	1494	1087	765	538	484	226	226	204

求平均价格 Y 关于使用年数 X 的回归方程。



作散点图

解：观察试验数据的散点图，y 与 x 呈指数关系，

设经验回归方程为

$$\boxed{y = ae^{bx}, \quad (b < 0)}$$

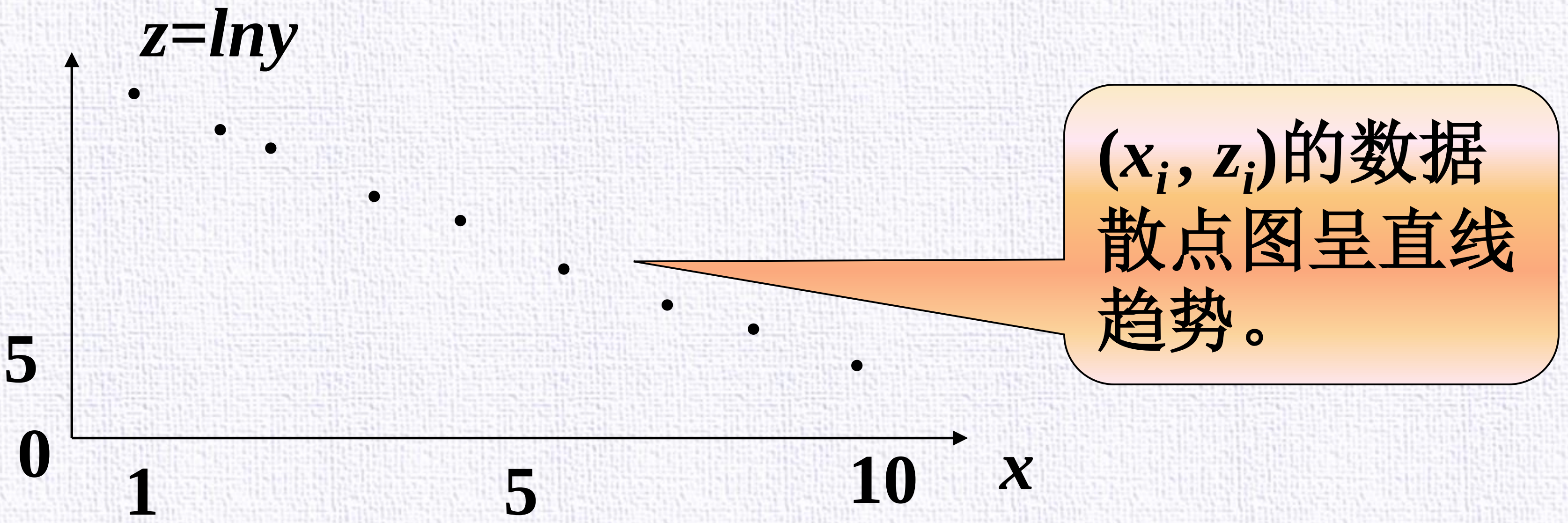
两边取对数，得 $\ln y = \ln a + bx$

令 $z = \ln y$, $x = x$, 记 $a' = \ln a$

经变换得回归方程为 $\boxed{z = a' + bx}$

记 $z_i = \ln y_i$，将原数据转换为（x_i, z_i），$i = 1, 2, \ldots, 10$.

x_i	1	2	3	4	5	6	7	8	9	10
z_i	7.88	7.57	7.31	6.99	6.64	6.29	6.18	5.67	5.42	5.32



$$\bar{x} = 5.5, \quad \bar{z} = 6.527 \qquad l_{xz} = \sum_{i=1}^{10} x_i z_i - 10\bar{x}\,\bar{z} \approx -24.5538$$

$$l_{xx} = \sum_{i=1}^{10} x_i^2 - 10(\bar{x})^2 = 38.5 - 10 \times 5.5^2 = 82.5$$

$$\hat{b} = \frac{l_{xz}}{l_{xx}} = -\frac{24.5538}{82.5} = -0.2976$$

$$\hat{a}' = \bar{z} - \hat{b}\bar{x} = 6.527 + 0.2976 \times 5.5 = 8.1642$$

从而得线性经验回归方程 $\hat{z} = 8.1642 - 0.2976x$

代入原变量，得非线性经验回归方程为

$$\hat{y} = e^{\hat{a}' + \hat{b}x} = e^{\hat{a}'} e^{\hat{b}x} = 3512.91 e^{-0.2976x}$$

检验X与Y是否存在显著的指数相关关系

⇕

检验X与$Z=\ln Y$的线性相关关系是否显著

有 $R = \dfrac{l_{xz}}{\sqrt{l_{xx}}\sqrt{l_{zz}}} = -0.996$，$|R| = 0.996 > 0.765 = R_{0.01}(8)$

可认为 X 与 Y 存在显著的指数相关关系。

$$\left[6-\frac{0.6}{3}\times 1.96, 6+\frac{0.6}{3}\times 1.96\right]=[5.608, 6.392]$$

2）需估计μ，σ^2未知，故选枢轴变量：$T=\dfrac{\bar{X}-\mu}{S/\sqrt{n-1}}\sim t(n-1)$

令$P\{|T|\leq t_{\alpha/2}(n-1)\}=1-\alpha$，解得$\mu$的置信区间

$$\left[\bar{X}-\frac{S}{\sqrt{n-1}}t_{\alpha/2}(n-1), \bar{X}+\frac{S}{\sqrt{n-1}}t_{\alpha/2}(n-1)\right]$$

$$\alpha=0.05,\ n=9,\ \bar{x}=\frac{1}{9}\sum_{i=1}^{9}x_i=6,\ S=\sqrt{\frac{1}{9}\sum_{i=1}^{9}(x_i-\bar{x})^2}\approx 0.5416,\ t_{0.025}(8)=2.306,$$

得μ的置信度为 0.95 的置信区间为

$$\left[6-\frac{0.5416}{2\sqrt{2}}\times 2.306, 6+\frac{0.5416}{2\sqrt{2}}\times 2.306\right]=[5.558, 6.442]$$

10、某商店一种产品的月销售量服从正态分布$N(\mu,\sigma^2)$，随机抽取 7 个月的销售量观察：

注：书店购买的答案错误百出，还是需要自己多看书，多动笔

定理6.2.4 设$X_1, X_2, \dots, X_n$是正态总体 $X \sim N(\mu, \sigma^2)$ 的样本，$\overline{X}, S^2$分别是样本均值和样本方差，则

(1) $\overline{X}$与S^2相互独立

(2) $\overline{X} \sim N(\mu, \dfrac{\sigma^2}{n})$

(3) $\dfrac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$

(4) $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

(1) (3) 证明： 不妨设总体服从标准正态。

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}$$

$$= (X_1, \dots, X_n) \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

经计算，发现A的特征值为{1,1,...,1,0}，因此存在正交变换T使得$TAT^{-1}=Diag\{0,1,1,\ldots,1\}$

$$T=\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\times1}} & -\frac{1}{\sqrt{2\times1}} & 0 & \cdots & \\ \frac{1}{\sqrt{3\times2}} & \frac{1}{\sqrt{3\times2}} & -\frac{2}{\sqrt{3\times2}} & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n\times(n-1)}} & \frac{1}{\sqrt{n\times(n-1)}} & \cdots & \cdots & \frac{-(n-1)}{\sqrt{n\times(n-1)}} \end{pmatrix}$$

如果令$(Y_1,Y_2,\ldots,Y_n)^T=T(X_1,X_2,\ldots,X_n)^T$

$$(X_1,\ldots,X_n)\,A\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix}=(X_1,\ldots,X_n)\,T^{-1}DT\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix}=\sum_{i=2}^{n}Y_i^2$$

因此得到$(n-1)S^2=\sum_{i=2}^{n}Y_i^2$

另一方面$\sqrt{n}\overline{X}=Y_1$

利用多元正态分布的线性不变性可以证明$(Y_1,Y_2,...,Y_n)$服从多元标准正态，即

$$(Y_1,Y_2,...,Y_n)\sim N(0,I_n)$$

因此Y_1与$\sum_{i=2}^{n}Y_i^2$独立，进一步证得样本均值和样本方差独立。

例 统计量的分布(之一)

设 X_1，X_2，…，X_n 是来自正态总体 $X\sim N(\mu,\sigma^2)$ 的容量为 n 的样本，求下列统计量的概率分布：

1. $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$　　2. $Y=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$

3.设$n=5$,若$a(X_1-X_2)^2+b(2X_3-X_4-X_5)^2\sim\chi^2(k)$,

则a,b,k各为多少？

解 1. $X_i\sim N(\mu,\sigma^2)\Rightarrow\sum_{i=1}^{n}X_i\sim N(n\mu,n\sigma^2)$

故 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\sim N(\mu,\frac{\sigma^2}{n})$

注：类似题目注意分布的本质以及独立性要求

2. $$\frac{X_i-\mu}{\sigma}\sim N(0,1)\Rightarrow\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2(n)$$

故 $$Y=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\sim\chi^2(n)$$

3. $$X_1-X_2\sim N(0,2\sigma^2),\Rightarrow U=\frac{X_1-X_2}{\sqrt{2}\sigma}\sim N(0,1)$$

$$2X_3-X_4-X_5\sim N(0,6\sigma^2),\Rightarrow V=\frac{2X_3-X_4-X_5}{\sqrt{6}\sigma}\sim N(0,1)$$

又U,V相互独立

$$\Rightarrow U^2+V^2=\frac{(X_1-X_2)^2}{2\sigma^2}+\frac{(2X_3-X_4-X_5)^2}{6\sigma^2}\sim\chi^2(2)$$

$$a=\frac{1}{2\sigma^2},\quad b=\frac{1}{6\sigma^2},\quad k=2.$$

例 统计量的分布(之二)

设 X_1，X_2，…，X_{n+m} 是来自正态总体 $X \sim N(0,\sigma^2)$ 的样本，求下列统计量的概率分布：

1. $Y = \dfrac{1}{\sigma^2}\sum_{i=1}^{n+m} X_i^2$　　2. $Z = \sqrt{\dfrac{m}{n}}\sum_{i=1}^{n} X_i \Big/ \sqrt{\sum_{i=n+1}^{n+m} X_i^2}$　　3. $\dfrac{1}{Z^2}$

1. $\dfrac{X_i}{\sigma} \sim N(0,1)$ 且所有 $\dfrac{X_i}{\sigma}$ 相互独立 $(i = 1,2,\dots,n+m)$

故 $$Y = \frac{1}{\sigma^2}\sum_{i=1}^{n+m} X_i^2 = \sum_{i=1}^{n+m}\left(\frac{X_i}{\sigma}\right)^2 \sim \chi^2(n+m)$$

2. 因 $\sum_{i=1}^{n} X_i \sim N(0, n\sigma^2) \Rightarrow U = \sum_{i=1}^{n} X_i \Big/ \sqrt{n\sigma^2} \sim N(0,1)$

同时 $V = \sum_{i=n+1}^{n+m} \frac{X_i^2}{\sigma^2} \sim \chi^2(m)$, U与V相互独立

从而有 $Z = \sqrt{\frac{m}{n}} \sum_{i=1}^{n} X_i \Big/ \sqrt{\sum_{i=n+1}^{n+m} X_i^2} = \frac{U}{\sqrt{V/m}} \sim t(m)$

3. $U \sim N(0,1) \Rightarrow U^2 \sim \chi^2(1)$

$V \sim \chi^2(m)$, U与V相互独立

故 $\frac{1}{Z^2} = \frac{V/m}{U^2/1} \sim F(m,1)$