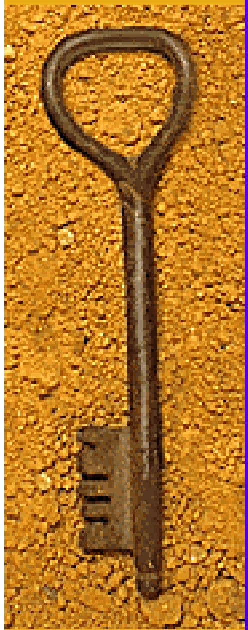


第三章 回归分析

处理变量与变量之间的统计相关关系

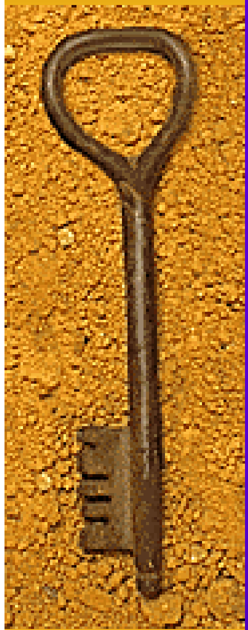
- 星系　氢含量、色指数、光度
- 太阳　耀斑、黑子、太阳射电辐射流量

统计相关关系　　不完全确定

观测误差 ↑　↓ 深入了解

函数关系　　完全确定

实质：概率统计＋最小二乘法



§ 一元线性回归

一 一元线性回归模型及参数估计

$$y_k = \beta_0 + \beta x_k + \varepsilon_k \qquad \text{一元线性回归模型}$$

$$E(y_k) = \beta_0 + \beta x_k \qquad \varepsilon_k \sim N(0, \sigma^2)$$

$$D(y_k) = \sigma^2 \qquad \text{正态误差回归模型}$$

寻找 β_0, β 的好的估计值，得到最能描述 y 和 x 关系的回归直线

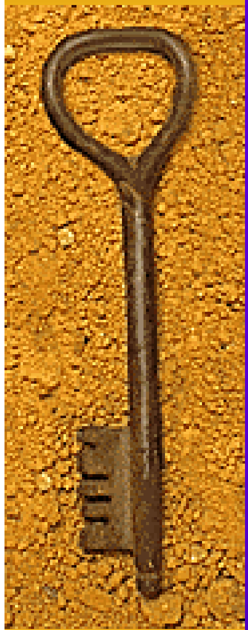
$$\hat{y}_k = b_0 + b x_k$$

利用最小二乘法给出 b_0, b 的计算公式

$$Q = \sum (y_k - \hat{y}_k)^2 = \sum (y_k - b_0 - b x_k)^2 = \min$$

$$\frac{\partial Q}{\partial b_0} = 0 \quad \rightarrow \quad b_0 = \frac{1}{n}\left(\sum y_k - b \sum x_k\right) = \bar{y} - b\bar{x}$$

$$\frac{\partial Q}{\partial b} = 0 \quad \rightarrow \quad b = \frac{\sum (x_k - \bar{x})(y_k - \bar{y})}{\sum (x_k - \bar{x})^2} = \frac{l_{xy}}{l_{xx}}$$



$$E(b_0)=\beta_0$$

$$E(b)=\beta$$

$$D(b_0)=\sigma^2[\frac{1}{n}+\frac{\overline{x}^2}{\sum(x_k-\overline{x})^2}]$$

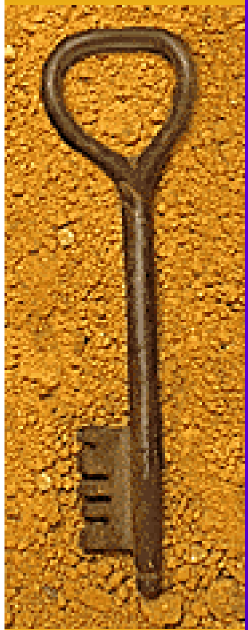
$$D(b)=\sigma^2[\frac{1}{\sum(x_k-\overline{x})^2}]$$

二 回归方程的显著性检验

$$\begin{aligned}\sum(y_k-\overline{y})^2&=\sum(y_k-\hat{y}_k+\hat{y}_k-\overline{y})^2\\&=\sum(y_k-\hat{y}_k)^2+\sum(\hat{y}_k-\overline{y})^2+2\sum(y_k-\hat{y}_k)(\hat{y}_k-\overline{y})\\&=\sum(y_k-\hat{y}_k)^2+\sum(\hat{y}_k-\overline{y})^2\\&=Q+U\end{aligned}$$

Q: 残差平方和　剩余平方和

U: 回归平方和　自变量变化引起



回归分析

1. 相关系数的检验

$$r^2 = U/l_{yy} \quad \Rightarrow \quad r = l_{xy}\Big/\sqrt{l_{xx}l_{yy}} \quad \Rightarrow \quad 0 \le |r| \le 1$$

$|r|$大　y与x线性相关密切

$|r|$小　y与x线性相关较弱

$r=1$　y与x完全线性相关

$r=0$　y与x毫无线性关系

$r>0$　$b>0$　正相关

$r<0$　$b<0$　负相关

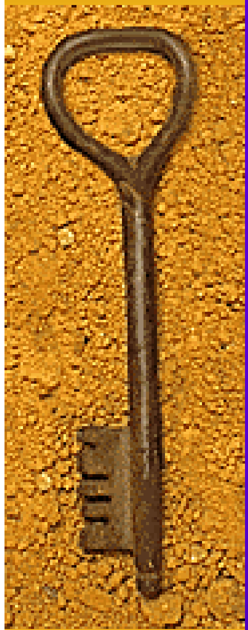
$r>r_\alpha$　r在α水平上显著

2. *F*检验（方差分析）

$$l_{yy}/\sigma^2 \sim \chi^2(n-1)$$

$$U/\sigma^2 \sim \chi^2(1)$$

$$Q/\sigma^2 \sim \chi^2(n-2)$$



回归分析

$$\frac{U(n-2)}{Q} \sim F(1, n-2)$$

$F > F_\alpha(1, n-2)$ 拒绝域 回归方程显著

相关系数显著性检验 $\Leftrightarrow$ 回归方程的F检验

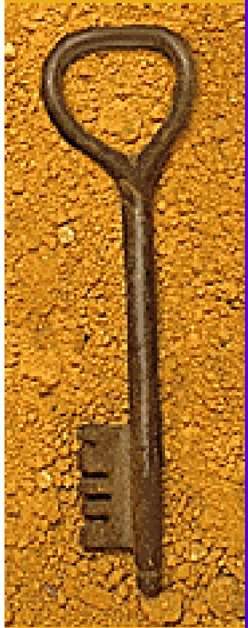
即 $r > r_\alpha \quad \Leftrightarrow \quad F > F_\alpha(1, n-2)$

证： $U = r^2 l_{yy} \quad Q = l_{yy} - U = (1 - r^2) l_{yy}$

$$F = \frac{U(n-2)}{Q} = \frac{(n-2)r^2}{1-r^2}$$

$$r = \sqrt{\frac{F}{(n-2)+F}}$$

$$r_\alpha = \sqrt{\frac{F_\alpha(1, n-2)}{(n-2)+F_\alpha(1, n-2)}}$$



三 回归系数和回归值的精度估计

β_0、β的区间估计

1. β的置信区间

1) σ已知

$$E(b)=\beta \quad D(b)=\sigma^2/l_{xx}$$

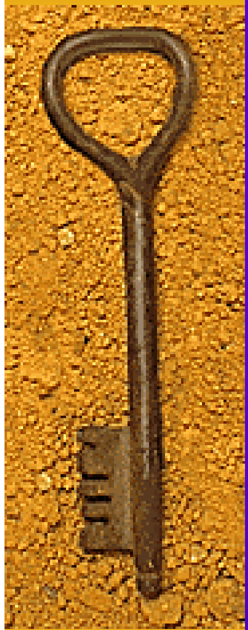
$$\Downarrow$$

$$b\sim N(\beta,\sigma^2/l_{xx})$$

$$\frac{b-\beta}{\sigma}\sqrt{l_{xx}}\sim N(0,1)$$

$$P(-u_\alpha<\frac{b-\beta}{\sigma}\sqrt{l_{xx}}<u_\alpha)=1-\alpha$$

β的区间估计 $(b-\mu_\alpha\sigma/\sqrt{l_{xx}},b+\mu_\alpha\sigma/\sqrt{l_{xx}})$



2) σ未知

$$S_y{}^2 = \hat{\sigma}^2 = Q/(n-2)$$

$$\frac{b-\beta}{S_y/\sqrt{l_{xx}}} \sim t(n-2)$$

$$\frac{b-\beta}{\sigma}\sqrt{l_{xx}} \sim N(0,1) \qquad \frac{Q}{\sigma^2} \sim \chi^2(n-2)$$

$$\Downarrow$$

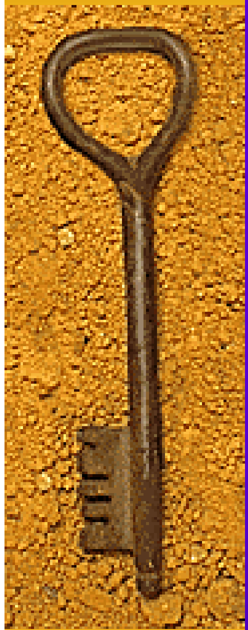
$$\frac{b-\beta}{\sigma}\sqrt{l_{xx}} \Bigg/ \sqrt{\frac{Q/\sigma^2}{n-2}} \sim t(n-2)$$

而 $S_y{}^2 = Q/n-2$

有 $\dfrac{b-\beta}{S_y/\sqrt{l_{xx}}} \sim t(n-2)$

$$P\left(-t_\alpha(n-2) < \frac{b-\beta}{S_y/\sqrt{l_{xx}}} < t_\alpha(n-2)\right) = 1-\alpha$$

β的区间估计 $\left(b - t_\alpha S_y/\sqrt{l_{xx}},\ b + t_\alpha S_y/\sqrt{l_{xx}}\right)$

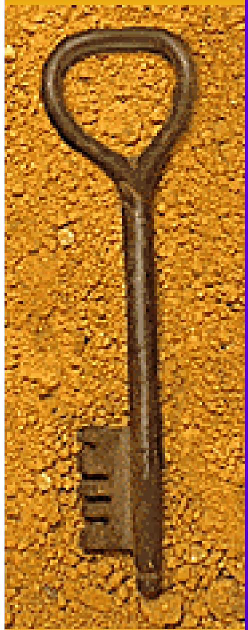


3. 回归值的置信区间

定义残差 $\delta_i = y_i - \hat{y}_i$

则 $E(\delta_i) = E(\beta_0 + \beta x_i + \varepsilon_i - b_0 - bx_i) = 0$

$$
\begin{aligned}
D(\delta_i) &= D(y_i - b_0 - bx_i) \\
&= D[y_i - \bar{y} - b(x_i - \bar{x})] \\
&= D\left[y_i - \bar{y} - \sum_k \frac{(x_k - \bar{x})(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} y_k \right] \\
&= D\left\{ y_i - \sum_k \left[\frac{1}{n} + \frac{(x_k - \bar{x})(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} \right] y_k \right\} \\
&= \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} \right] \sigma^2
\end{aligned}
$$



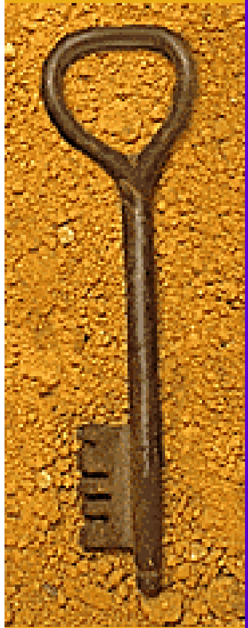
回归分析

$$\delta \sim N(0,\sigma\sqrt{1+\frac{1}{n}+\frac{(x_i-\overline{x})^2}{\sum_j (x_j-\overline{x})^2}})$$

$$P(-\delta_n < y-\hat{y} < \delta_n) = 1-\alpha$$

y的区间估计 $(y-\delta_n, \hat{y}+\delta_n)$

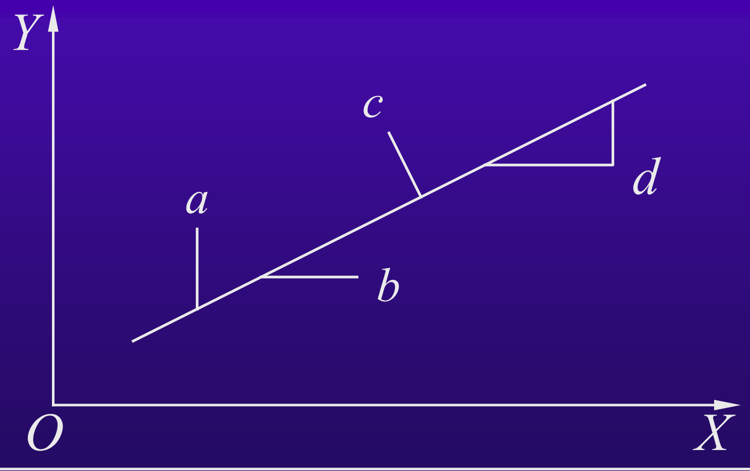
$$\delta_N = u_\alpha \sigma\sqrt{1+\frac{1}{n}+\frac{(x_i-\overline{x})^2}{\sum_j (x_j-\overline{x})^2}}$$

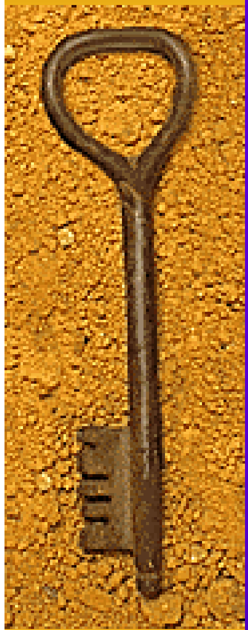


四 五种一元线性回归及其在天文上的应用

1. 五种线性回归方法

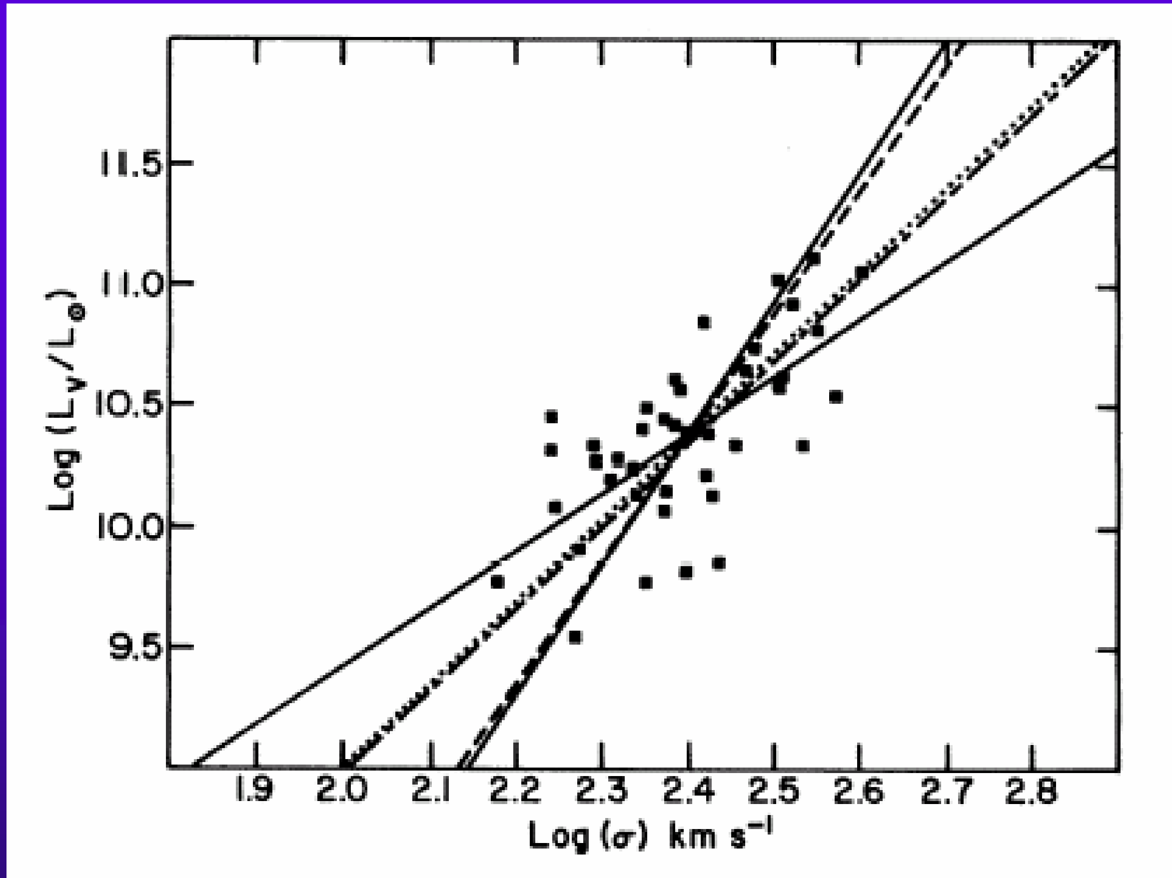
1) OLS($Y|X$)：观测点和回归直线上同一 x 的 y 的差；
2) 逆回归OLS($X|Y$)：观测点和回归直线上同一 y 的 x 的差；
3) 正交回归线OR ：观测点到回归线的垂直距离；
4) 简化主轴回归RMA ：观测点对回归线在垂直、水平两个方向测量的距离；
5) OLS平分线： OLS($Y|X$)和OLS($X|Y$)的平分线。



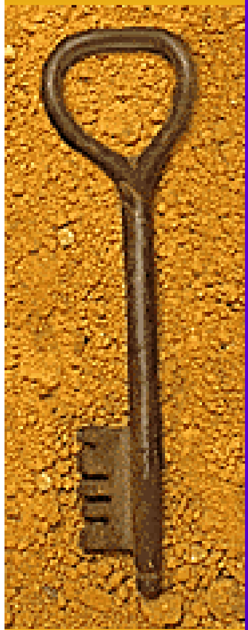


回归分析

应用五种回归方法测椭圆星系速度弥散 σ和光学光度之间的关系$L\sim\sigma^n$



图：L和 σ的对数散点图及它们的五种回归线：1. OLS(Y | X)
2. OLS(X | Y) 3. OLS平分线(点虚线) 4. OR(虚线) 5. RMA(点线)



§ 曲线回归分析

一 曲线回归类型的确定

1. 散点图

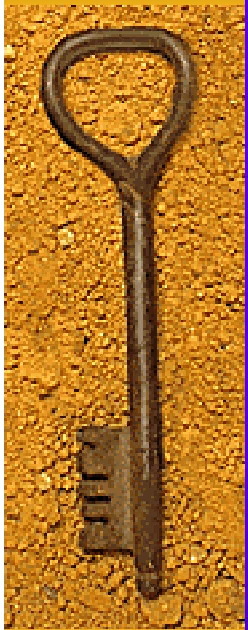
   利用观测数据的散点图，对比已知函数形式的各种曲线，选择最为接近的曲线作为回归函数

2. 多项式

   $$y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_m x^m+\varepsilon$$

二 曲线回归参数的确定

$$\mathrm{I}\quad \begin{cases} y=\beta_0+\beta\,\mathrm{e}^x \\ y=\beta_0+\beta\ln x \\ y=\beta_0+\beta x^l \end{cases} \quad\Rightarrow\quad y=\beta_0+\beta x' \quad \begin{array}{l} x'=\mathrm{e}^x \\ x'=\ln x \\ x'=x^l \end{array}$$



$$\text{II}\quad \begin{cases} y=\dfrac{1}{\beta_0+\beta\,\mathrm{e}^x} \\ y=\beta_0\,\mathrm{e}^{\beta x} \\ y=\beta_0 x^{\beta} \end{cases} \quad\Rightarrow\quad y'=\beta_0{}'+\beta x' \quad \begin{matrix} y'=1/y \\ \begin{cases}\beta_0{}'=\ln\beta_0 \\ y'=\ln y\end{cases} \\ \begin{cases}\beta_0{}'=\ln\beta_0 \\ x'=\ln x \\ y'=\ln y\end{cases} \end{matrix}$$

$$\text{III}\quad y=\mathrm{e}^{\beta_1 x}+\mathrm{e}^{\beta_2 x}$$

I、II进行变换，转化为线性回归；III泰勒级数展开，变为线性。

三　曲线回归的有效性检验

相关指数　$R=\sqrt{1-\dfrac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}}$

标准剩余差　$S_y=\sqrt{\dfrac{\sum(y_i-\hat{y}_i)^2}{n-2}}$