



数学实验

Experiments in Mathematics

实验11 回归分析

Dept. of Mathematical Sciences
Tsinghua University, Beijing 100084, China

1

回归分析(Regression Analysis)

- 1、实例和基本概念
- 2、多元线性回归
- 3、MATLAB统计工具箱(Statistics Toolbox)的使用

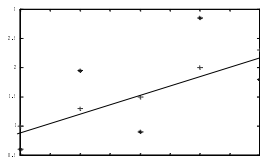
2

拟合问题简例

$x = [0 \ 1 \ 2 \ 3 \ 4]$, $y = [1.0 \ 1.3 \ 1.5 \ 2.0 \ 2.3]$ (+号)
 $x = [0 \ 1 \ 2 \ 3 \ 4]$, $z = [0.6 \ 1.95 \ 0.9 \ 2.85 \ 1.8]$ (*号)

直线拟合:
 $a = \text{polyfit}(x, y, 1)$,
 $b = \text{polyfit}(x, z, 1)$,

得到
 $a = 0.33 \ 0.96$
 $b = 0.33 \ 0.96$



同一条直线 $y = 0.33x + 0.96$ ($z = 0.33x + 0.96$)

问题: 你相信哪个拟合结果? 怎样给以定量评价。

3

回归分析的主要任务

对拟合问题作统计分析, 给出可信程度的定量评价

回归分析在一组数据的基础上研究以下问题:

- (1) 建立因变量 y 与自变量 x_1, x_2, \dots, x_m 之间的回归模型;
- (2) 对回归模型的可信度进行检验;
- (3) 判断每个自变量 x_i ($i=1, \dots, m$) 对 y 的影响是否显著;
- (4) 诊断回归模型是否适合这组数据;
- (5) 利用回归模型对 y 进行预报或控制。

4

例1 年龄与运动能力

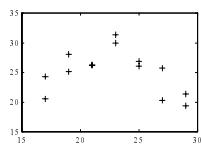
将 17 至 29 岁的运动员每两岁一组分为 7 组, 每组两人测量其旋转定向能力, 以考察年龄对这种运动能力的影响。

年龄	17	19	21	23	25	27	29
第一人	20.48	25.13	26.15	30.0	26.1	20.3	19.35
第二人	24.35	28.11	26.3	31.4	26.92	25.7	21.3

建立运动能力与年龄的关系

将数据(年龄 x , 运动能力 y)作图

数据散点图显示: y 与 x 呈非线性关系, 可建立二次(或高次)多项式回归模型。



5

例2 商品销售量与价格

某厂生产的一种电器的销售量 y 与竞争对手的价格 x_1 (元) 和本厂的价格 x_2 (元) 有关。

下表是该商品在 10 个城市的销售记录。

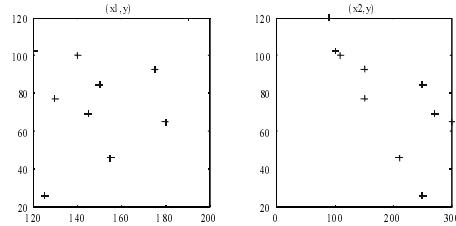
x_1	120	140	190	130	155	175	125	145	180	150
x_2	100	110	90	150	210	150	250	270	300	250
Y	102	100	120	77	46	93	26	69	65	85

- 1) 试根据这些数据建立 y 与 x_1 和 x_2 的关系式, 对得到的模型和系数进行检验。
- 2) 若某市本厂产品售价 160 元, 竞争对手售价 170 元, 预测该市的销售量。

6

例2 商品销售量与价格

将 $(x_1, y), (x_2, y)$ 各10个点分别画图



y 与 x_2 有较明显的线性关系, y 与 x_1 之间的关系难以确定

需要对模型 $y=f(x_1, x_2)$ 作几种尝试, 用统计分析决定优劣。

7

多元线性回归

一元线性回归 $y = \beta_0 + \beta_1 x$

多元线性回归 $y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m \ (m \geq 2)$

多元线性回归的一般形式 $y = \beta_0 + \beta_1 f_1(x) + \dots + \beta_m f_m(x)$

$x = (x_1, \dots, x_l) \sim$ 多元变量 $f_j \ (j = 1, \dots, m) \sim$ 已知函数

\mathbf{y} 对回归系数 $\beta = (\beta_0, \beta_1, \dots, \beta_m)$ 线性

$y = \beta_0 + \beta_1 f_1(x) + \dots + \beta_m f_m(x) \iff y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$

变量代换

多元线性回归
标准形

8

多元线性回归

y 的主要部分: $\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$

y 的随机误差: $\varepsilon \sim N(0, \sigma^2)$

模型

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m + \varepsilon \\ \varepsilon \sim N(0, \sigma^2), \sigma \text{未知, 与} x \text{无关} \end{cases}$$

得到 n 个独立数据 $(y_i, x_{i1}, \dots, x_{im}), i = 1, \dots, n, n > m,$

y_i 的随机误差为 ε_i

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, n \end{cases}$$

由数据估计参数 $\beta_0, \beta_1, \dots, \beta_m$, 使得误差最小。

9

多元线性回归

容量为 n 的样本数据 $(y_i, x_{i1}, \dots, x_{im}), i = 1, \dots, n, n > m,$

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \beta = [\beta_0, \beta_1, \dots, \beta_m]^T$$

模型的
矩阵形式

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 I) \end{cases}$$

10

参数估计

$Y = X\beta + \varepsilon$ 用最小二乘法估计参数 β

误差平方和 $Q(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta)$

求 β 使 $Q(\beta)$ 最小

$$\frac{dQ}{d\beta} = -2X^T(Y - X\beta)$$

$$\frac{dQ}{d\beta} = 0 \iff \hat{\beta} = (X^T X)^{-1} X^T Y$$

思考: X 具有什么性质 $X^T X$ 才可逆; 实际问题中怎样保证 X 具有这种性质; $n > m$ 的要求为什么是必要的?

11

参数估计

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m)$$

将 $\hat{\beta}$ 代入原模型 $y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$

得到 y 的估计值 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_m x_m$

数据 Y 的拟合值 $\hat{Y} = X\hat{\beta}$

拟合残差 $e = Y - \hat{Y} \sim$ 随机误差 ε 的估计

残差平方和 (剩余平方和): $Q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

12

统计分析 $\hat{\beta} = (X^T X)^{-1} X^T Y$ $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

1) $\hat{\beta}$ 是 β 的线性无偏最小方差估计

$\hat{\beta}$ 对 Y 线性 $E(\hat{\beta}) = \beta$ 在线性无偏估计中, $\hat{\beta}$ 的方差最小

2) 正态分布 $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$

3) 残差平方和 Q $EQ = (n-m-1)\sigma^2$, $Q/\sigma^2 \sim \chi^2(n-m-1)$

$s^2 = \frac{Q}{n-m-1} = \hat{\sigma}^2$ s^2 (剩余方差) 是 σ^2 的无偏估计

4) 分解 Y 的样本方差 $S = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

Q -残差平方和 (ε的影响) U -回归平方和 (x的影响) $= Q + U$

模型的假设检验

多元线性回归 $y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$ $U = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

y 与 x_1, x_2, \dots, x_m 之间是否存在线性关系? $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

假设检验 $H_0: \beta_j = 0$ ($j=1, \dots, m$) $U/\sigma^2 \sim \chi^2(m)$

H_0 成立, $F = \frac{U/m}{Q/(n-m-1)} \sim F(m, n-m-1)$ $Q/\sigma^2 \sim \chi^2(n-m-1)$

显著性水平 α 的检验规则:
若 $F < F_{1-\alpha}(m, n-m-1)$, 接受 H_0 ; 否则, 拒绝 H_0

拒绝 H_0 只说明线性关系的不显著

$R^2 = U/S$, $R \in [0, 1]$ ~ 相关系数 R 越大, y 与 x_1, x_2, \dots, x_m 的关系越密切

系数的假设检验 $\hat{\beta} = (X^T X)^{-1} X^T Y$

$H_0: \beta_j = 0$ ($j=1, \dots, m$) 被拒绝时, β_j 不全为 0, 应作如下检验

第 j 个系数假设检验 $H_0^{(j)}: \beta_j = 0$ ($j=1, \dots, m$)

$H_0^{(j)}$ 成立时 $t_j = \frac{\hat{\beta}_j / \sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)}} \sim t(n-m-1)$ $s^2 = \frac{Q}{n-m-1} = \hat{\sigma}^2$

$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$, c_{jj} 是 $(X^T X)^{-1}$ 对角线上的元素 显著性水平 α

检验规则 若 $|t_j| < t_{1-\alpha/2}(n-m-1)$, 接受 $H_0^{(j)}$, 否则, 拒绝。

β_j 的置信区间 $[\hat{\beta}_j - t_{1-\alpha/2}(n-m-1) s \sqrt{c_{jj}}, \hat{\beta}_j + t_{1-\alpha/2}(n-m-1) s \sqrt{c_{jj}}]$

给定 α, m, n , s, c_{jj} 越小, β_j 的估计精度越高。

多元线性回归的 MATLAB 实现

基本命令 `b=regress(Y,X)`
`[b,bint,r,rint,stats]=regress(Y,X,alpha)`
`rcoplot(r,rint)`

输入 $Y = (y_1, \dots, y_n)^T$, $X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}$ **alpha**: 显著水平

输出 **b**: 回归系数估计值, **bint**: 回归系数置信区间
r: 残差向量, **rint**: 残差向量置信区间

stats: (**s1**, **s2**, **s3**), **s1**: 相关系数 R^2 , **s2**: 模型检验 F 值, **s3**: F 分布随机变量取值大于 **s2** 的概率。 **s3** < α 时拒绝 H_0 。

`rcoplot(r,rint)`: 残差及置信区间绘图

例 商品销售量与价格 回归模型 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

`x1=[120 140 190 130 155 175 125 145 180 150];`
`x2=[100 110 90 150 210 150 250 270 300 250];`
`y=[102 100 120 77 46 93 26 69 65 85];`
`x=[ones(10,1) x1' x2'];`
`[b,bint,r,rint,stats]=regress(y,x);`
`rcoplot(r,rint);`

演示 regress example.m

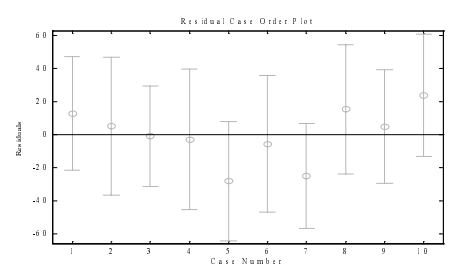
结果: **b** = 66.5176 0.4139 -0.2698
bint = -32.5060 165.5411
-0.2018 1.0296
-0.4611 -0.0785
stats = 0.6527 6.5786 0.0247

分析: $\alpha=0.05$, 模型可用; $\alpha=0.01$, 不能用。 $R^2=0.6527$, 线性相关性较小。 $\hat{\beta}_0, \hat{\beta}_1$ 置信区间包含零点。

模型改进

例 商品销售量与价格 回归模型 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

rcoplot(r,rint): 残差及置信区间绘图



残差置信区间均包含零点, 表明无异常数据。

例 合金的强度 y (kg/mm²) 与碳含量 x (%) hejin.m

x	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.20	0.21	0.23
y	420	415	450	455	450	475	490	550	500	550	555	605

模型 $y = \beta_0 + \beta_1 x$

b = 27.0269 140.6194
 bint = 22.3226 31.7313
 111.7842 169.4546
 stats = 0.9219 118.0670 0.0000

第8个残差置信区间不含零点，剔除重算。

b = 27.0992 137.8085
 bint = 23.8563 30.3421
 117.8534 157.7636
 stats = 0.9644 244.0571 0.0000

R², F 明显变大，应该用此结果。

多元线性回归----多项式回归

一元多项式模型 $y = a_m x^m + \dots + a_1 x + a_0$

MATLAB命令 `[p,S]=polyfit(x,y,m)`

输入: x ~自变量(数据, 行向量), y ~因变量(同前), m ~多项式阶数;
 输出: $p=(a_m, \dots, a_1, a_0)$ ~多项式系数, S ~多项式结构;

[Y,delta]=polyconf(p,x,S,alpha)

输入: x (同上), p, S (polyfit的输出), α ~显著性水平;
 输出: $Y \sim y$ 的拟合值, $\delta \sim Y \pm \delta$ 是 y 的置信区间
 (α 缺省时为0.05)

20

例 年龄与运动能力 演示 onearpoly.m

选用2次模型 $y = a_2 x^2 + a_1 x + a_0$

程序

```

x1=17:2:29;x=[x1,x1];
y=[20.48 25.13 26.15 30.0 26.1 20.3 19.35...
24.35 28.11 26.3 31.4 26.92 25.7 21.3];
[p,S]=polyfit(x,y,2);p
[Y,delta]=polyconf(p,x,S);Y
y1=mean(y);
rsquare=sum((Y-y1).^2)./sum((y-y1).^2),
s=sqrt(sum((y-Y).^2)./11),

```

结果 $p = -0.2003 \quad 8.9782 \quad -72.2150 = (a_2, a_1, a_0)$
 $Y = 22.5243 \quad 26.0582 \quad \dots \quad 19.6904$ 拟合值(可用于作图)
 R^2, s (衡量优劣)

MATLAB命令 `polytool(x,y,m)`

输入: x, y, m (同前);
 输出: 拟合及置信区间曲线绘图(交互式画面)。

例 年龄与运动能力 `polytool(x,y,2)`

多元线性回归----多元二项式回归

适用问题: 建立因变量与多变量的二次关系

$y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$

$y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m + \sum_{j=1}^m \beta_{jj} x_j^2$

$y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m + \sum_{1 \leq j,k \leq m} \beta_{jk} x_j x_k$

$y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m + \sum_{1 \leq j,k \leq m} \beta_{jk} x_j x_k$

MATLAB命令: `rstool(x,y,'model',alpha)`

x --- $n \times m$ 矩阵, y --- n 维列向量, $model$ ---上面4个模型之一
 α ---显著性水平, 缺省时为0.05。

例 商品的销售量与价格

模型 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + 2\beta_{12} x_1 x_2 + \beta_{22} x_2^2$

$x1=[120 \ 140 \ 190 \ 130 \ 155 \ 175 \ 125 \ 145 \ 180 \ 150];$
 $x2=[100 \ 110 \ 90 \ 150 \ 210 \ 150 \ 250 \ 270 \ 300 \ 250];$
 $y=[102 \ 100 \ 120 \ 77 \ 46 \ 93 \ 26 \ 69 \ 65 \ 85];$
 $x=[x1' \ x2'];$

演示 `multivarqua.m`

`rstool(x,y,'quadratic')`

rstool有两类输出

1) Export~向工作区传送参数: **beta**--回归系数, **rmse**--剩余标准差, **residuals**--残差(向量);

```
beta = -307.3600 7.2032 -1.7374 0.0001 -0.0226 0.0037
rmse = 18.6064
residuals = 6.5136 -12.6257 0.0280 6.3717 -19.7331 7.8895
           -11.2896 5.6333 -5.2394 22.4518
```

2) Model~ 在上述4种模型(线性、纯2次、纯交互、完全2次)中选择。

25

例 商品的销售量与价格

以剩余标准差rmse最小为标准, 比较4种模型

model=linear: rmse=18.7362

model=purequadratic: rmse=16.6436

model=interaction: rmse=19.1626

model=quadratic: rmse=18.6064

最终模型 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$

26

逐步回归

- 从众多影响因变量的因素中选出影响显著的自变量建立回归模型;
- 从便于应用的角度, 自变量应尽量少;
- 从候选自变量集合 $S = \{x_1, \dots, x_m\}$ 中选出一子集 S_1 (含 $l \leq m$ 个自变量) 与因变量 y 构造回归模型, 其优劣由剩余标准差 s 度量;

$$s^2 = Q / (n - l - 1), n \sim \text{数据量}, Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

当影响显著的自变量进入模型时, Q 明显下降, s 减小; 而影响很小的自变量进入模型时, Q 下降不大, l 的增加会使 s 变大。

逐步回归的基本思路

1. 选自变量初始子集 S_0 (相互独立性较强);
2. 确定引入水平 α_{in} , 剔除水平 α_{out} ;
3. 从 S_0 外引入对影响最大的 x (要满足 α_{in}), 再剔除影响最小的 x (要满足 α_{out}), 建立模型, 得新子集 S_1 ;
4. 重复3, 直到在 $\alpha_{in}, \alpha_{out}$ 下没有引入和剔除为止。

28

逐步回归的MATLAB实现

命令: **stepwise(x,y)**

输入: $x \sim n \times m$ 阵(自变量), $y \sim n \times 1$ 阵(因变量)

输出: 3个交互式画面:

- 1) **stepwise table** (回归系数及其置信区间数值, 模型统计量 **RMSE**, **Rsquare**, **F**, **p**);
 - 2) **stepwise history** (**RMSE** 及其置信区间);
 - 3) **stepwise plot** (显示回归系数及其置信区间, 可向工作区传送)。
- 由2) 可调回原有的模型。

29

逐步回归的MATLAB实现

演示:
stepwisel.m

例 水泥凝固时放出的热量 y 与水泥中 4 种化学成分 x_1, x_2, x_3, x_4 有关, 今测得一组数据, 试用逐步回归来确定一个线性模型。

初始模型

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_4 x_4$$

序号	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

- 命令: **stepwise(x,y,inmodel,alpha)**
- 可将任何 x_i 用已知函数变换后, 再作逐步回归。

30

非线性回归

y 对参数 $\beta_0, \beta_1 \dots \beta_m$ 非线性

MATLAB命令:

`nlinfit(x,y,'model',beta0)`~输出回归系数等,
模型由model.m文件给出;

`nlintool(x,y,'model',beta0)`~输出交互式画面;
`nlparci,nlpredci`等。

参看MATLAB帮助系统, 及书294页。

31

布置作业

目的

- 1、了解回归分析的基本原理
- 2、根据问题的要求提出模型
- 3、对已经确定的模型, 确定参数、使用MATLAB

作业

1), 3), 7)

1. 目的。
2. 内容(对每一题): 模型(对应用题); 算法设计; 计算结果; 结果分析; 附程序(必要时加说明语句)。
3. 收获和建议。

32