数学建模的基本方法

机理分析　测试分析

由于客观事物内部规律的复杂及人们认识程度的限制，无法分析实际对象内在的因果关系，建立合乎机理规律的数学模型.

通过对数据的统计分析，找出与数据拟合最好的模型.

回归模型是用统计分析方法建立的最常用的一类模型.

- 不涉及回归分析的数学原理和方法 .
- 通过实例讨论如何选择不同类型的模型 .
- 对软件得到的结果进行分析，对模型进行改进.

10.1 牙膏的销售量

问题 建立牙膏销售量与价格、广告投入之间的模型；

预测在不同价格和广告费用下的牙膏销售量.

收集了30个销售周期本公司牙膏销售量、价格、广告费用，及同期其他厂家同类牙膏的平均售价 .

销售周期	本公司价格(元)	其他厂家价格(元)	广告费用(百万元)	价格差（元）	销售量(百万支)
1	3.85	3.80	5.50	-0.05	7.38
2	3.75	4.00	6.75	0.25	8.51
…	…	…	…	…	…
29	3.80	3.85	5.80	0.05	7.93
30	3.70	4.25	6.80	0.55	9.26

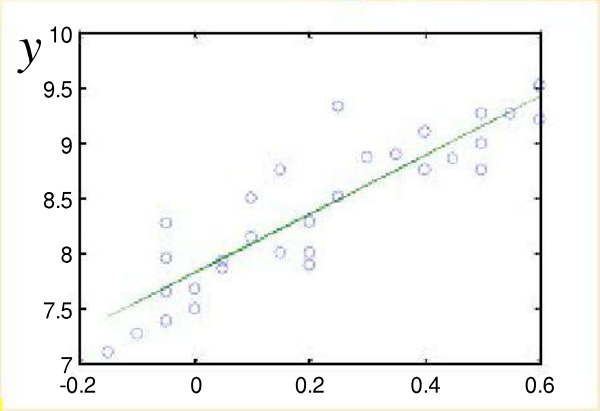
基本模型

y ~公司牙膏销售量

x_1~其他厂家与本公司价格差

x_2~公司广告费用

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$



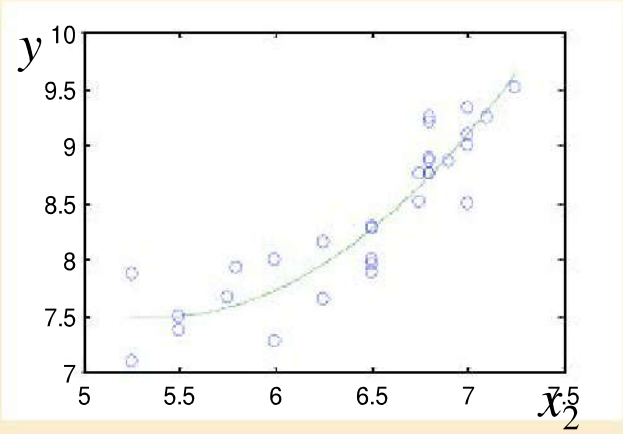
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

y~被解释变量（因变量）

x_1, x_2~解释变量(回归变量, 自变量)

$\beta_0, \beta_1, \beta_2, \beta_3$ ~回归系数

ε~随机误差（均值为零的正态分布随机变量）



$$y = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \varepsilon$$

模型求解

MATLAB 统计工具箱

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$　由数据 y, x_1, x_2 估计 β

```
[b,bint,r,rint,stats]=regress(y,x,alpha)
```

输入

y~n维数据向量

x= $[1 \quad x_1 \quad x_2 \quad x_2^2]$ ~n×4数据矩阵, 第1列为全1向量

alpha(置信水平,0.05)

输出

b~β的估计值

bint~b的置信区间

r ~残差向量y-xb

rint~r的置信区间

参数	参数估计值	置信区间
β_0	17.3244	[5.7282　28.9206]
β_1	1.3070	[0.6829　1.9311]
β_2	-3.6956	[-7.4989　0.1077]
β_3	0.3486	[0.0379　0.6594]
R^2=0.9054	F=82.9409	p<0.0001　s^2=0.0490

Stats~
检验统计量
R^2, F, p, s^2

结果分析

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$

参数	参数估计值	置信区间
β_0	17.3244	[5.7282 28.9206]
β_1	1.3070	[0.6829 1.9311]
β_2	-3.6956	[-7.4989 0.1077]
β_3	0.3486	[0.0379 0.6594]
R^2=0.9054	F=82.9409	p<0.0001 s^2=0.0490

y的90.54%可由模型确定

F值远超过F检验的临界值

p值远小于α=0.05

模型从整体上看成立

β_2的置信区间包含零点
(右端点距零点很近)

x_2对因变量y 的
影响不太显著

x_2^2项显著

可将x_2保留在模型中

销售量预测

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2$$

价格差x_1=其他厂家价格x_3-本公司价格x_4

估计x_3 调整x_4 ⇨ 控制x_1 ⇨ 通过x_1, x_2预测y

控制价格差x_1=0.2元，投入广告费x_2=6.5百万元

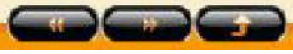
$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2=8.2933\ (\text{百万支})$$

销售量预测区间为 [7.8230，8.7636]（置信度95%）

上限用作库存管理的目标值　下限用来把握公司的现金流

若估计x_3=3.9，设定x_4=3.7，则可以95%的把握

知道销售额在 7.8230×3.7≈ 29（百万元）以上



模型改进

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$

x 和x_2对y的影响独立

参数	参数估计值	置信区间
β_0	17.3244	[5.7282 28.9206]
β_1	1.3070	[0.6829 1.9311]

x_1和x_2对y的影响有交互作用

两模型销售量预测比较

控制价格差x_1=0.2元，投入广告费x_2=6.5百万元

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2$$

预测值 $\hat{y}=8.2933$　　预测区间 [7.8230，8.7636]

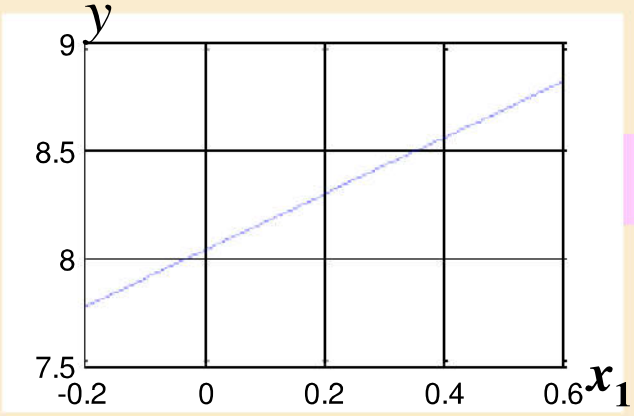
$$\hat{y}=\beta_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2+\hat{\beta}_4x_1x_2$$

预测值 $\hat{y}=8.3272$　　预测区间 [7.8953，8.7592]

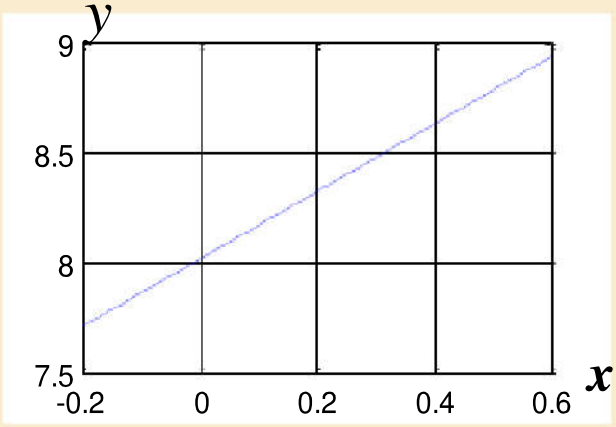
$\hat{y}$ 略有增加　　预测区间长度更短

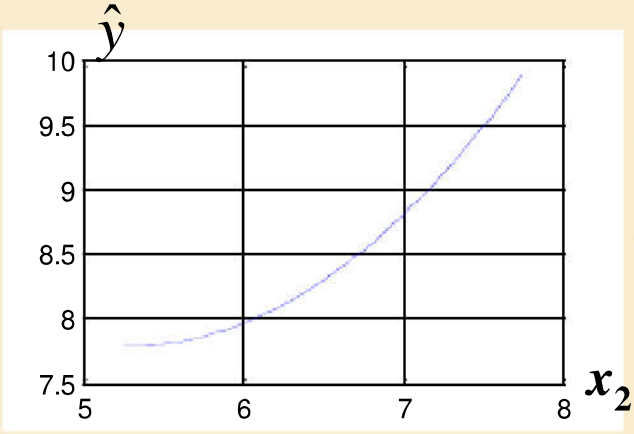
两模型 $\hat{y}$ 与x_1, x_2关系的比较

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 \qquad \hat{y} = \beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 x_2$$

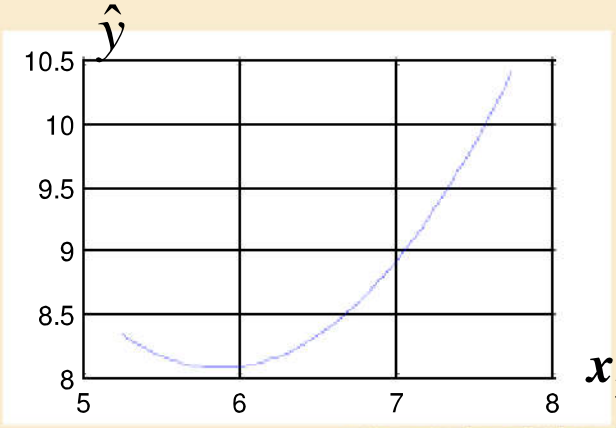


x_2=6.5





x_1=0.2



交互作用影响的讨论

$$\hat{y}=\beta_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2+\hat{\beta}_4x_1x_2$$

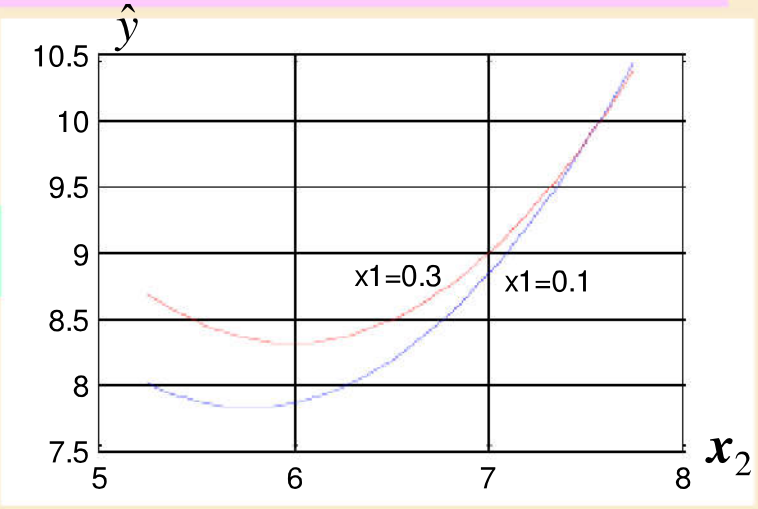
价格差 x_1=0.1

$$\hat{y}\big|_{x_1=0.1}=30.2267-7.7558x_2+0.6712x_2^2$$

价格差 x_1=0.3

$$\hat{y}\big|_{x_1=0.3}=32.4535-8.0513x_2+0.6712x_2^2$$

$x_2<7.5357$ ⇨ $\hat{y}\big|_{x_1=0.3}>\hat{y}\big|_{x_1=0.1}$



⇨ 价格优势会使销售量增加

加大广告投入使销售量增加

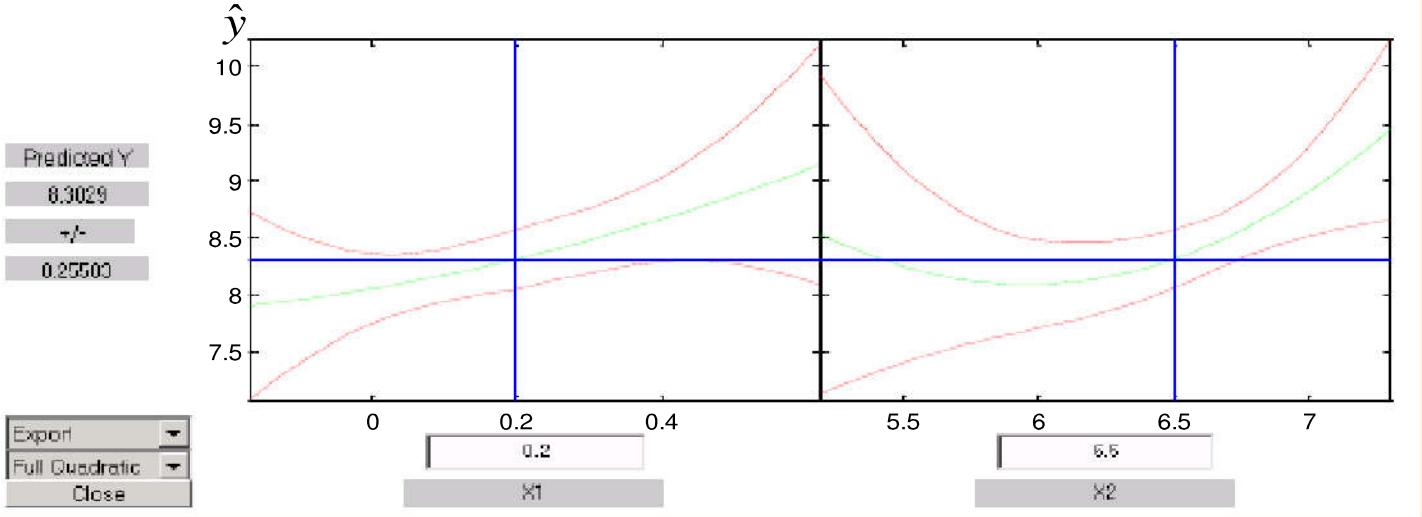
（x_2大于6百万元）

价格差较小时增加的速率更大 ⇨ 价格差较小时更需要靠广告来吸引顾客的眼球

完全二次多项式模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$

MATLAB中有命令rstool直接求解



从输出 Export 可得 $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$

鼠标移动十字线(或下方窗口输入)可改变x_1, x_2,
左边窗口显示预测值 $\hat{y}$ 及预测区间

牙膏的销售量

建立统计回归模型的基本步骤

- 根据已知数据从常识和经验分析, 辅之以作图, 决定回归变量及函数形式(先取尽量简单的形式).
- 用软件(如MATLAB统计工具箱)求解.
- 对结果作统计分析: R^2,F, p, s^2是对模型整体的评价, 回归系数置信区间是否含零点，用于检验回归变量对因变量的影响是否显著.
- 模型改进, 如增添二次项、交互项等.
- 对因变量进行预测.

10.2 软件开发人员的薪金

建立模型研究薪金与资历、管理责任、教育程度的关系.

分析人事策略的合理性，作为新聘用人员薪金的参考.

46名软件开发人员的档案资料

编号	薪金	资历	管理	教育	编号	薪金	资历	管理	教育
01	13876	1	1	1	42	27837	16	1	2
02	11608	1	0	3	43	18838	16	0	2
03	18701	1	1	3	44	17483	16	0	1
04	11283	1	0	2	45	19207	17	0	2
…	…	…	…	…	46	19346	20	0	1

资历~ 从事专业工作的年数；管理~ 1=管理人员, 0=非管理人员；
教育~ 1=中学，2=大学，3=更高程度.

分析与假设

y~ 薪金，x_1~资历（年）

$x_2=1$~ 管理人员，$x_2=0$~ 非管理人员

教育

1=中学

2=大学

3=更高

$$x_3=\begin{cases}1, & \text{中学}\\ 0, & \text{其他}\end{cases}$$

$$x_4=\begin{cases}1, & \text{大学}\\ 0, & \text{其他}\end{cases}$$

中学：$x_3=1, x_4=0$；

大学：$x_3=0, x_4=1$；

更高：$x_3=0, x_4=0$

假设

- 资历每加一年, 薪金的增长是常数；
- 管理、教育、资历之间无交互作用.

线性回归模型

$$y=a_0+a_1x_1+a_2x_2+a_3x_3+a_4x_4+\varepsilon$$

$a_0, a_1, \ldots, a_4$是待估计的回归系数，ε是随机误差

模型求解

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \varepsilon$$

参数	参数估计值	置信区间
a_0	11033	[10258 11807]
a_1	546	[484 608]
a_2	6883	[6248 7517]
a_3	-2994	[-3826 -2162]
a_4	148	[-636 931]
R^2=0.9567	F=226	p<0.0001 s^2=10^6

R^2,F, p→ 模型整体上可用

x_1~资历(年)

x_2=1~ 管理，

x_2=0~ 非管理

中学：x_3=1, x_4=0;

大学：x_3=0, x_4=1;

更高：x_3=0, x_4=0.

资历增加1年薪金增长546

管理人员薪金多6883

中学程度薪金比更高的少2994

大学程度薪金比更高的多148

a_4置信区间包含零点，解释不可靠!

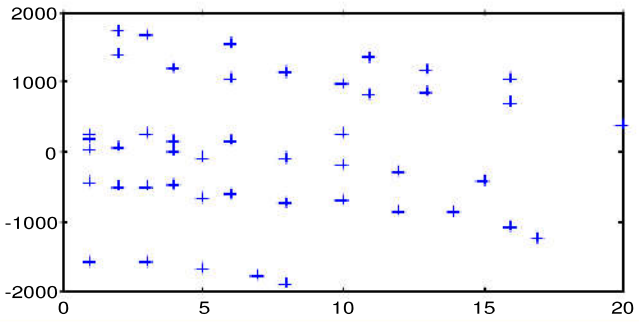
结果分析

残差分析方法

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \hat{a}_3 x_3 + \hat{a}_4 x$$

残差 $e = y - \hat{y}$

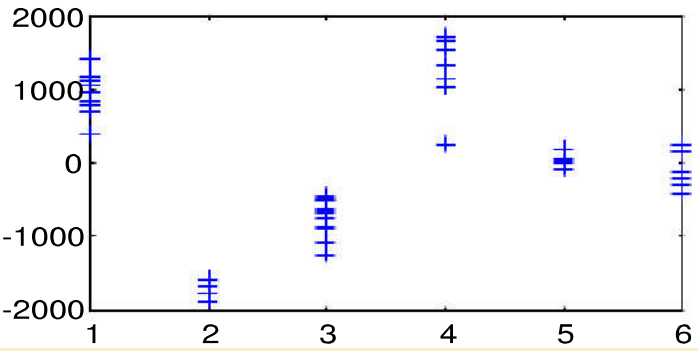
e 与资历x_1的关系



残差大概分成3个水平，6种管理—教育组合混在一起，未正确反映.

管理与教育的组合

组合	1	2	3	4	5	6
管理	0	1	0	1	0	1
教育	1	1	2	2	3	3

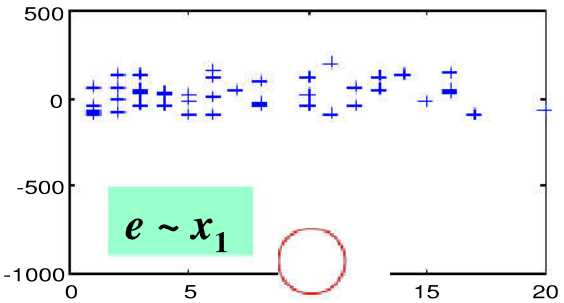


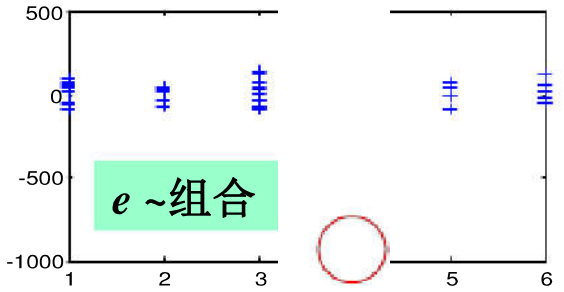
残差全为正，或全为负，管理—教育组合处理不当.

应在模型中增加管理x_2与教育x_3, x_4的交互项 .

进一步的模型　增加管理x_2与教育x_3, x_4的交互项

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_2x_3 + a_6x_2x_4 + \varepsilon$$



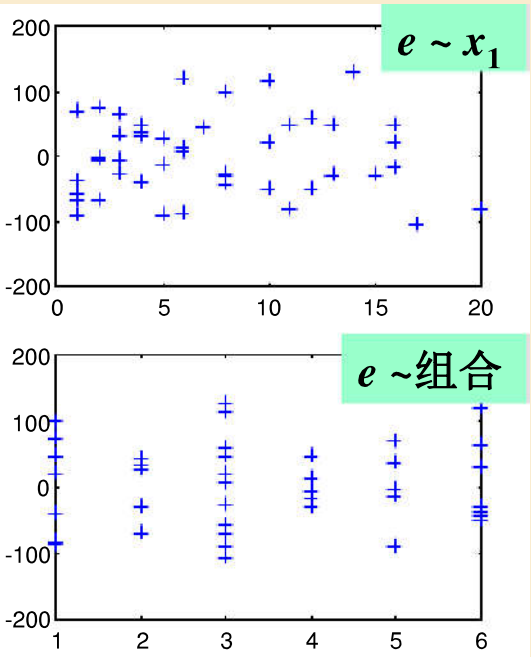


R^2, F有改进, 所有回归系数置信区间不含零点, 模型完全可用

消除了不正常现象

异常数据(33号)应去掉!

去掉异常数据后的结果



R^2: 0.9567 →0.9988→0.9998

F： 226 → 554 → 36701

s^2： 10^4 → 3×10^4 → 4×10^3

残差图十分正常

最终模型的结果可以应用

置信区间长度更短

模型应用

$$\hat{y}=\hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2+\hat{a}_3x_3+\hat{a}_4x_4+\hat{a}_5x_2x_3+\hat{a}_6x_2x_4$$

制订6种管理—教育组合人员的“基础”薪金(资历为0)

$x_1=0$； $x_2=1$~ 管理，$x_2=0$~ 非管理

中学：$x_3=1, x_4=0$ ； 大学：$x_3=0, x_4=1$； 更高：$x_3=0, x_4=0$

组合	管理	教育	系数	“基础”薪金
1	0	1	a_0+a_3	9463
2	1	1	$a_0+a_2+a_3+a_5$	13448
3	0	2	a_0+a_4	10844
4	1	2	$a_0+a_2+a_4+a_6$	19882
5	0	3	a_0	11200
6	1	3	a_0+a_2	18241

大学程度管理人员比更高程度管理人员的薪金高.

大学程度非管理人员比更高程度非管理人员的薪金略低.

软件开发人员的薪金

对定性因素(如管理、教育)可以引入0-1变量处理，0-1变量的个数可比定性因素的水平少1.

残差分析方法可以发现模型的缺陷，引入交互作用项常常能够改善模型.

剔除异常数据，有助于得到更好的结果.

注：可以直接对6种管理—教育组合引入5个0-1变量.

10.3 酶促反应

问题 研究酶促反应（酶催化反应）中嘌呤霉素对反应速度与底物（反应物）浓度之间关系的影响.

建立数学模型，反映该酶促反应的速度与底物浓度以及经嘌呤霉素处理与否之间的关系.

方案 设计了两个实验：酶经过嘌呤霉素处理；酶未经嘌呤霉素处理. 实验数据见下表.

底物浓度(ppm)		0.02		0.06		0.11		0.22		0.56		1.10	
反应速度	处理	76	47	97	107	123	139	159	152	191	201	207	200
	未处理	67	51	84	86	98	115	131	124	144	158	160	/

酶促反应的基本性质

底物浓度较小时，反应速度大致与浓度成正比；

底物浓度很大、渐进饱和时，反应速度趋于固定值.

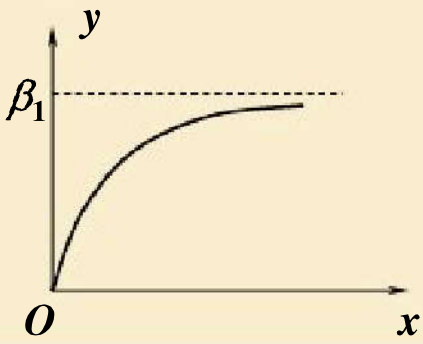
基本模型

Michaelis-Menten模型

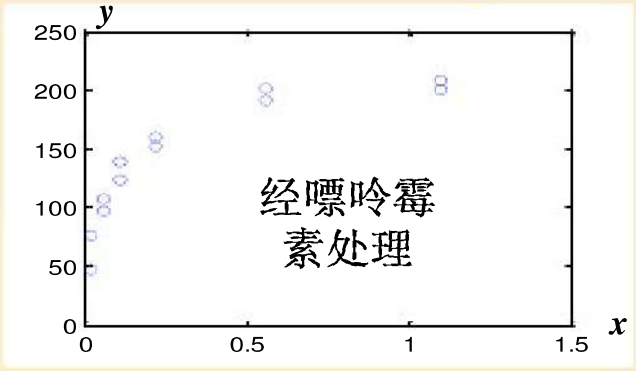
y ~ 酶促反应的速度，x ~底物浓度

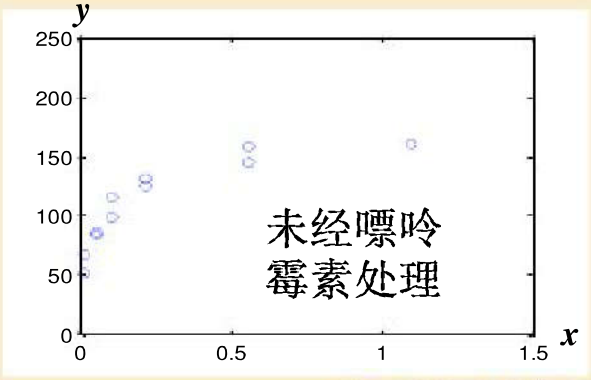
$$y = f(x,\beta) = \frac{\beta_1 x}{\beta_2 + x}$$

β_1, β_2 ~ 待定系数



实验数据





线性化模型

$$y=\frac{\beta_1 x}{\beta_2+x} \quad \Rightarrow \quad \frac{1}{y}=\frac{1}{\beta_1}+\frac{\beta_2}{\beta_1}\frac{1}{x}=\theta_1+\theta_2\frac{1}{x}$$

对β_1, β_2非线性 ⇒ 对θ_1, θ_2线性

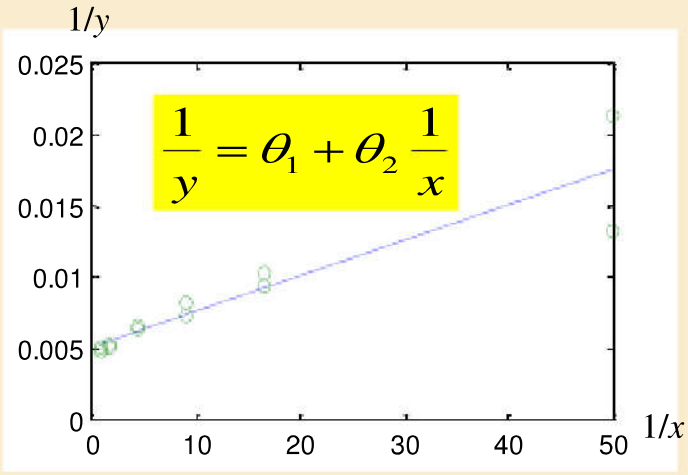
经嘌呤霉素处理后实验数据的估计结果

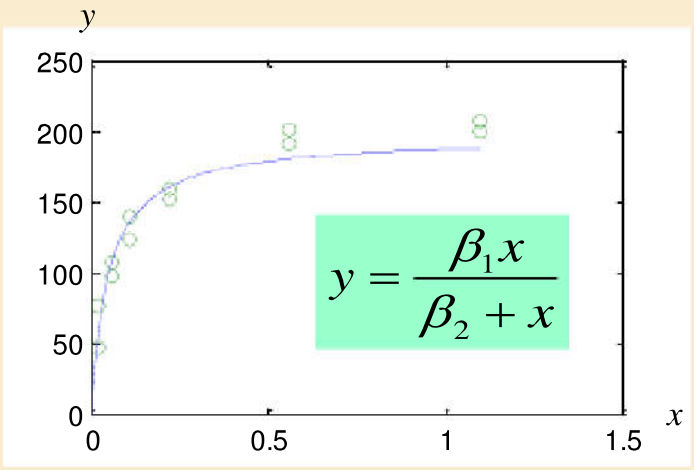
参数	参数估计值（$\times 10^{-3}$）	置信区间（$\times 10^{-3}$）
θ_1	5.1072	[3.5386 6.6758]
θ_2	0.2472	[0.1757 0.3188]
R^2=0.8557 F=59.2975 p<0.0001 s^2=3.5806 $\times 10^{-6}$		

$\hat{\beta}_1=1/\hat{\theta}_1=195.8027$

$\hat{\beta}_2=\hat{\theta}_2/\hat{\theta}_1=0.04841$

线性化模型结果分析





1/x较小时有很好的线性趋势，1/x较大时出现很大的起落.

x较大时，y有较大偏差

- **参数估计时，x较小(1/x很大)的数据控制了回归参数的确定.**

非线性模型参数估计

MATLAB 统计工具箱

[beta,R,J] = nlinfit (x,y,'model',beta0)

输入

x~自变量数据矩阵

y ~因变量数据向量

$$y = \frac{\beta_1 x}{\beta_2 + x}$$

beta0~线性化模型估计结果

model ~模型的函数M文件名

beta0 ~给定的参数初值

输出

beta ~参数的估计值

R ~残差，J ~估计预测误差的Jacobi矩阵

beta的置信区间

betaci =nlparci(beta,R,J)

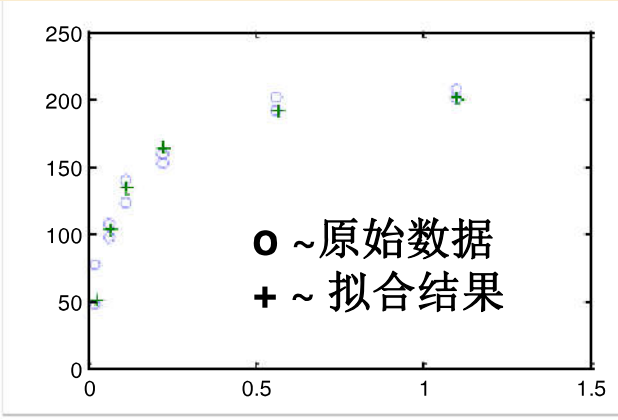
```
x=                    ; y=                    ;
beta0=[195.8027  0.04841];
[beta,R,J]=nlinfit(x,y,'f1',beta0);
betaci=nlparci(beta,R,J);
beta, betaci
```

```
function  y=f1(beta, x)
y=beta(1)*x./(beta(2)+x);
```

非线性模型结果分析

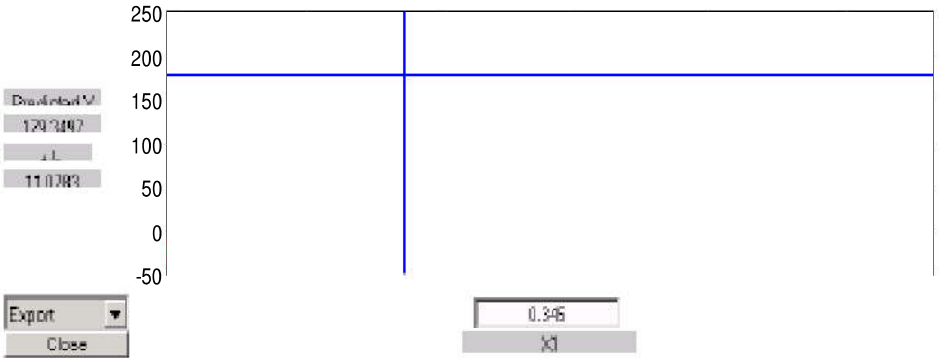
$$y = \frac{\beta_1 x}{\beta_2 + x}$$

参数	参数估计值	置信区间



半速度点(达到最终速度一半时的x值)为 $\hat{\beta}_2 = 0.0641$

命令nlintool 给出交互画面



混合反应模型

在同一模型中考虑嘌呤霉素处理的影响

$$y=\frac{\beta_1 x}{\beta_2+x} \quad \Rightarrow \quad y=\frac{(\beta_1+\gamma_1 x_2)\,x_1}{(\beta_2+\gamma_2 x_2)+x_1}$$

x_1为底物浓度， x_2为一示性变量

x_2=1表示经过处理，x_2=0表示未经处理

β_1是未经处理的最终反应速度

γ_1是经处理后最终反应速度的增长值

β_2是未经处理的反应的半速度点

γ_2是经处理后反应的半速度点的增长值

混合模型求解

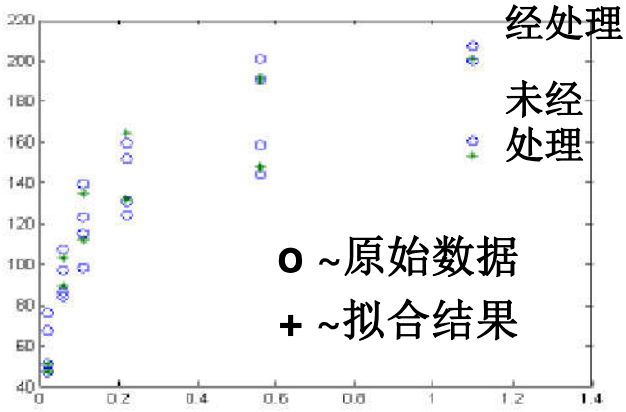
$$y=\frac{(\beta_1+\gamma_1 x_2)\ x_1}{(\beta_2+\gamma_2 x_2)+x_1}$$

用nlinfit 和 nlintool命令

参数初值(基于对数据的分析) $\beta_1^0=170,\ \gamma_1^0=60,\ \beta_2^0=0.05,\ \gamma_2^0=0.01$

估计结果和预测

参数	参数估计值	置信区间
β_1	160.2802	[145.8466 174.7137]
β_2	0.0477	[0.0304 0.0650]
γ_1	52.4035	[32.4130 72.3941]
γ_2	0.0164	[-0.0075 0.0403]



剩余标准差s=10.4000

γ_2置信区间包含零点，表明γ_2对因变量y的影响不显著

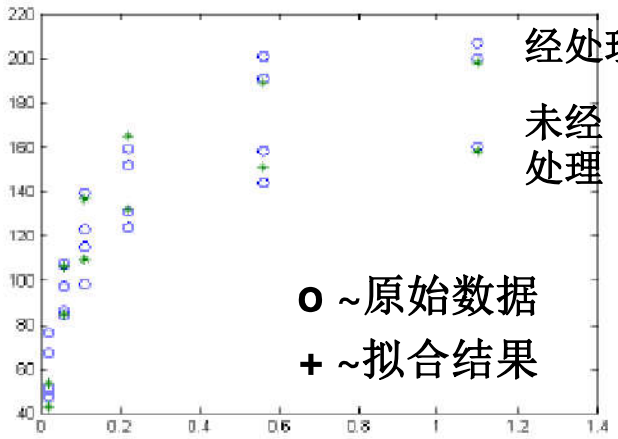
⇨ 经嘌呤霉素处理的作用不影响半速度点参数

简化的混合模型

$$y=\frac{(\beta_1+\gamma_1 x_2)\,x_1}{(\beta_2+\gamma_2 x_2)+x_1} \Rightarrow y=\frac{(\beta_1+\gamma_1 x_2)\,x_1}{\beta_2+x_1}$$

估计结果和预测

参数	参数估计值	置信区间
β_1	166.6025	[154.4886 178.7164]
β_2	0.0580	[0.0456 0.0703]
γ_1	42.0252	[28.9419 55.1085]



简化的混合模型形式简单，参数置信区间不含零点.

剩余标准差 $s = 10.5851$，比一般混合模型略大.

一般混合模型与简化混合模型预测比较

$$y=\frac{(\beta_1+\gamma_1 x_2)\,x_1}{(\beta_2+\gamma_2 x_2)+x_1}$$

$$y=\frac{(\beta_1+\gamma_1 x_2)\,x_1}{\beta_2+x_1}$$

预测区间为
预测值 ± Δ

实际值	一般模型预测值	Δ(一般模型)	简化模型预测值	Δ(简化模型)
67	47.3443	9.2078	42.7358	5.4446
51	47.3443	9.2078	42.7358	5.4446
84	89.2856	9.5710	84.7356	7.0478
…	…	…	…	…
191	190.8329	9.1484	189.0574	8.8438
201	190.8329	9.1484	189.0574	8.8438
207	200.9688	11.0447	198.1837	10.1812
200	200.9688	11.0447	198.1837	10.1812

简化混合模型的预测区间较短，更为实用、有效.

酶促反应

机理分析

反应速度与底物浓度的关系 ⇨ 非线性关系

求解线性模型 ⇨ 求解非线性模型

发现问题，
得参数初值

嘌呤霉素处理对反应速度与底物浓度关系的影响

⇨ 混合模型 ⇨ 简化模型

引入0-1变量

检查参数置信区间
是否包含零点

注：非线性模型拟合程度的评价无法直接利用线性模型的方法，但R^2 与s仍然有效.

10.4 投资额与生产总值和物价指数

问题

建立投资额模型，研究某地区实际投资额与国民生产总值（GNP）及物价指数（PI）的关系.

根据对未来GNP及PI的估计，预测未来投资额.

该地区连续20年的统计数据

年份序号	投资额	国民生产总值	物价指数	年份序号	投资额	国民生产总值	物价指数
1	90.9	596.7	0.7167	11	229.8	1326.4	1.0575
2	97.4	637.7	0.7277	12	228.7	1434.2	1.1508
3	113.5	691.1	0.7436	13	206.1	1549.2	1.2579
4	125.7	756.0	0.7676	14	257.9	1718.0	1.3234
5	122.8	799.0	0.7906	15	324.1	1918.3	1.4005
6	133.3	873.4	0.8254	16	386.6	2163.9	1.5042
7	149.3	944.0	0.8679	17	423.0	2417.8	1.6342
8	144.2	992.7	0.9145	18	401.9	2631.7	1.7842
9	166.4	1077.6	0.9601	19	474.9	2954.7	1.9514
10	195.0	1185.9	1.0000	20	424.5	3073.0	2.0688

投资额与国民生产总值和物价指数

分析 许多经济数据在时间上有一定的滞后性.

以时间为序的数据，称为时间序列.

时间序列中同一变量的顺序观测值之间存在自相关.

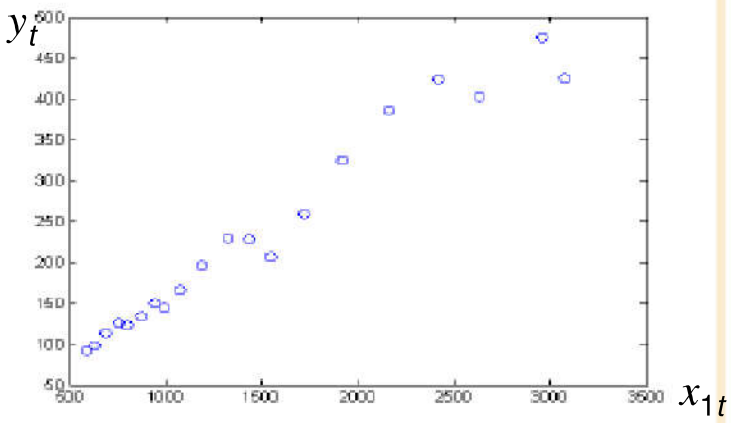
若采用普通回归模型直接处理，将会出现不良后果.

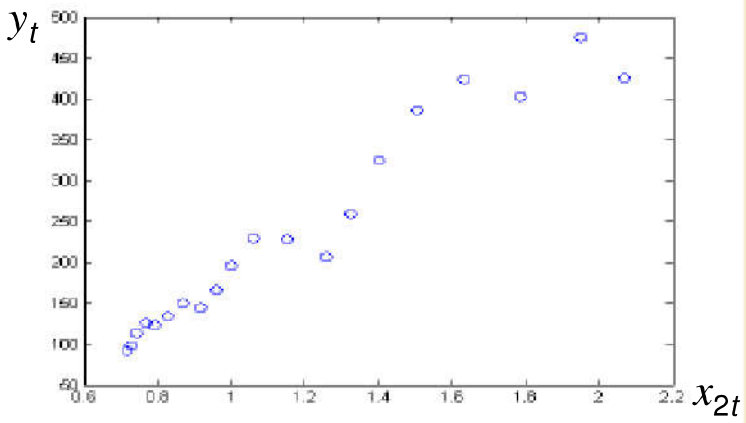
需要诊断并消除数据的自相关性，建立新的模型.

年份序号	投资额	国民生产总值	物价指数	年份序号	投资额	国民生产总值	物价指数
1	90.9	596.7	0.7167	11	229.8	1326.4	1.0575
2	97.4	637.7	0.7277	12	228.7	1434.2	1.1508
3	113.5	691.1	0.7436	13	206.1	1549.2	1.2579
4	125.7	756.0	0.7676	14	257.9	1718.0	1.3234
…	…	…	…	…	…	…	…

基本回归模型

t ~年份， y_t ~ 投资额，x_{1t}~ GNP, x_{2t} ~ 物价指数





投资额与 GNP及物价指数间均有很强的线性关系

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$$

β_0, β_1, β_2 ~回归系数

ε_t ~对t相互独立的零均值正态随机变量

基本回归模型的结果与分析

MATLAB 统计工具箱

参数	参数估计值	置信区间
β_0	322.7250	[224.3386 421.1114]
β_1	0.6185	[0.4773 0.7596]
β_2	-859.4790	[-1121.4757 -597.4823]
R^2= 0.9908	F= 919.8529	p<0.0001 s^2=161.7

$$\hat{y}_t = 322.725 + 0.6185\, x_{1t} - 859.479\, x_{2t}$$

剩余标准差
s=12.7164

模型优点 R^2=0.9908，拟合度高

模型缺点 没有考虑时间序列数据的滞后性影响.
可能忽视了随机误差存在自相关；如果存在自相关性，用此模型会有不良后果.

自相关性的定性诊断

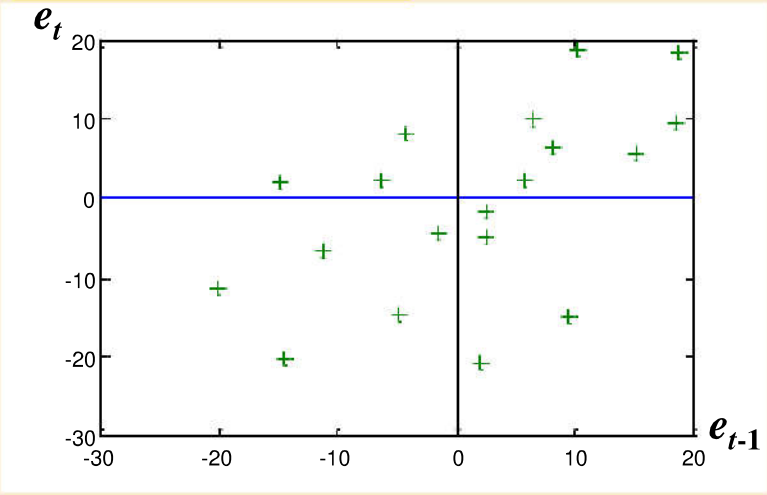
残差诊断法

模型残差 $e_t = y_t - \hat{y}_t$

e_t为随机误差ε_t的估计值

在MATLAB工作区中输出

作残差 e_t~e_{t-1} 散点图



大部分点落在第1, 3象限 ⇨ ε_t 存在正的自相关

大部分点落在第2, 4象限 ⇨ ε_t 存在负的自相关

自相关性直观判断 ⇨ 基本回归模型的随机误差项ε_t 存在正的自相关

自回归性的定量诊断

D-W检验

自回归模型 $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t$

$\beta_0, \beta_1, \beta_2$ ~回归系数　　ρ~自相关系数　　$|\rho| \le 1$

u_t ~对t相互独立的零均值正态随机变量

$\rho = 0$	⇨	无自相关性
$\rho > 0$	⇨	存在正自相关性
$\rho < 0$	⇨	存在负自相关性

如何估计ρ ⇦ D-W统计量

如何消除自相关性 ⇦ 广义差分法

D-W统计量与D-W检验

$$\hat{\rho}=\sum_{t=2}^{n}e_te_{t-1}\Big/\sum_{t=2}^{n}e_t^{\,2}$$

$$DW=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=2}^{n}e_t^2}\approx 2\left[1-\frac{\sum_{t=2}^{n}e_te_{t-1}}{\sum_{t=2}^{n}e_t^2}\right]=2(1-\hat{\rho})$$

n较大

$-1\le\hat{\rho}\le 1\rightarrow 0\le DW\le 4$

$\hat{\rho}=1\rightarrow DW=0$

$\hat{\rho}=-1\rightarrow DW=4$

$\hat{\rho}=0\rightarrow DW=2$

检验水平, 样本容量, 回归变量数目

D-W分布表

检验临界值d_L和d_U

由DW值的大小确定自相关性

广义差分变换 $DW=2(1-\hat{\rho})$ ⇨ $\hat{\rho}=1-\frac{DW}{2}$

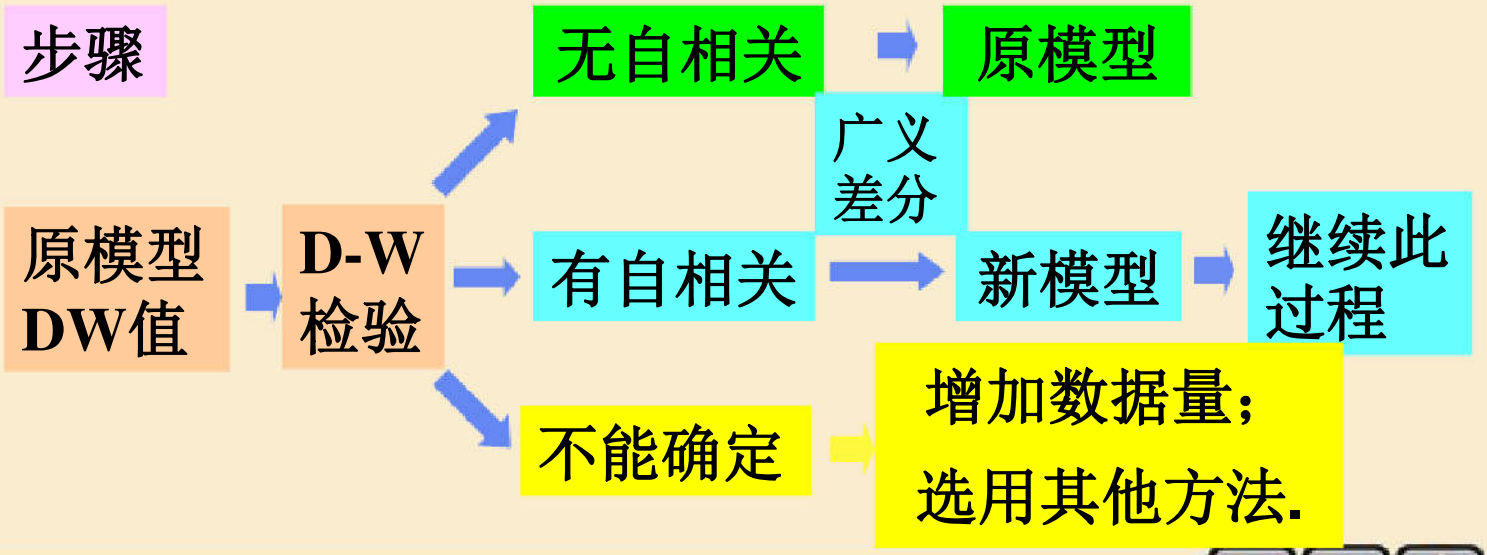
原模型 $y_t=\beta_0+\beta_1x_{1t}+\beta_2x_{2t}+\varepsilon_t,\quad \varepsilon_t=\rho\varepsilon_{t-1}+u_t$

变换 $y_t^*=y_t-\rho y_{t-1},\ x_{it}^*=x_{it}-\rho x_{i,t-1},\quad i=1,2$

新模型 $y_t^*=\beta_0^*+\beta_1x_{1t}^*+\beta_2x_{2t}^*+u_t\quad \beta_0^*=\beta_0(1-\rho)$

以$\beta_0^*,\beta_1,\beta_2$为回归系数的普通回归模型

步骤



投资额新模型的建立

$$DW=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=2}^{n}e_t^2}$$

原模型　　DW_{old}=

残差e_t　　0.8754

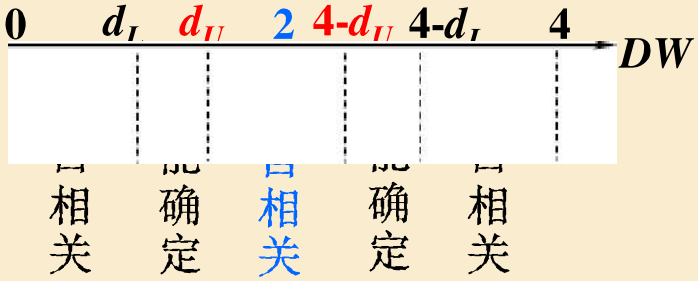
样本容量n=20, 回归变量数目k=3, α=0.05

查表

临界值d_L=1.10, d_U=1.54

$DW_{old}<d_L$ ⇒ 原模型有正自相关

$$\hat{\rho}=1-DW/2=0.5623$$



作变换

$$y_t^*=y_t-0.5623\,y_{t-1}$$

$$x_{it}^*=x_{it}-0.5623x_{i,t-1},\quad i=1,2$$

投资额新模型的建立

$$y_t^* = y_t - 0.5623\, y_{t-1} \qquad x_{it}^* = x_{it} - 0.5623 x_{i,t-1}, \quad i = 1,2$$

$$y_t^* = \beta_0^* + \beta_1 x_{1t}^* + \beta_2 x_{2t}^* + u_t$$

由数据 $y_t^*, x_{1t}^*, x_{2t}^*$ 估计系数 β_0^*, β_1, β_2

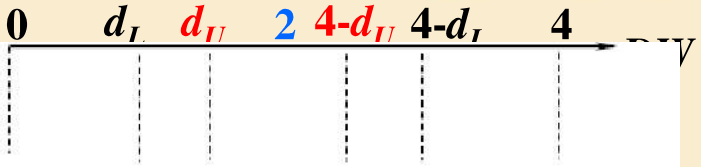
参数	参数估计值	置信区间
β_0^*	163.4905	[1265.4592 2005.2178]
β_1	0.6990	[0.5751 0.8247]
β_2	-1009.0333	[-1235.9392 -782.1274]
R^2= 0.9772	F=342.8988	p<0.0001 s^2=96.58

总体效果良好

剩余标准差

s_{new}= 9.8277 < s_{old}=12.7164

新模型的自相关性检验



新模型 $\hat{y}_t^* = 163.4905 + 0.699\, x_{1t}^* - 1009.033\, x_{2t}^*$

还原为原始变量

$$\hat{y}_t = 163.4905 + 0.5623 y_{t-1} + 0.699 x_{1,t} - 0.3930 x_{1,t-1} - 1009.0333 x_{2,t} + 567.3794 x_{2,t-1}$$

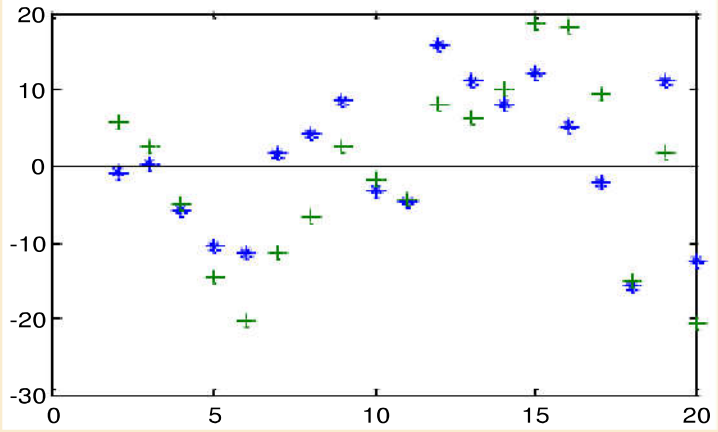
一阶自回归模型

模型结果比较

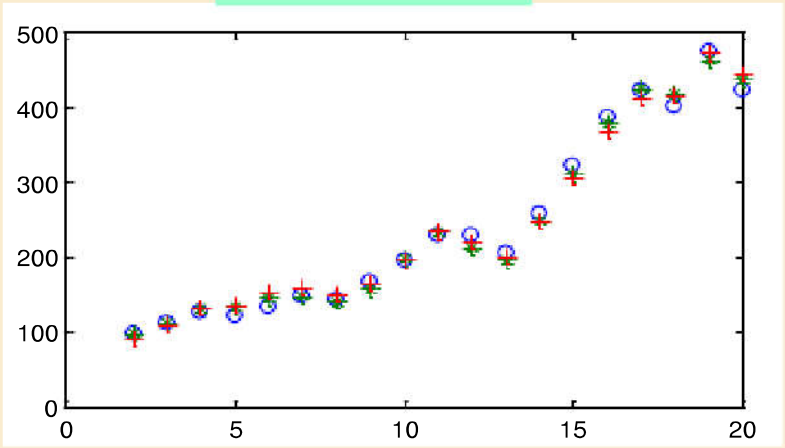
基本回归模型 $\hat{y}_t = 322.725 + 0.6185\, x_{1t} - 859.479\, x_{2t}$

一阶自回归模型 $\hat{y}_t = 163.4905 + 0.5623 y_{t-1} + 0.699 x_{1,t} - 0.3930 x_{1,t-1}$
$- 1009.0333 x_{2,t} + 567.3794 x_{2,t-1}$

残差图比较



拟合图比较



新模型 e_t~ *，原模型 e_t~ +

新模型 $\hat{y}_t$ ~ *，新模型 $\hat{y}_t$ ~ +

一阶自回归模型残差e_t比基本回归模型要小.

$\hat{y}_t$ 较小是由于y_{t-1}=424.5过小所致

10.5 教学评估

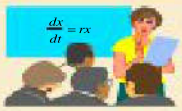
问题 为了考评教师的教学质量，教学研究部门对学生进行问卷调查，得到15门课程各项评分的平均值.

编号	X_1	X_2	X_3	X_4	X_5	X_6	Y
201	4.46	4.42	4.23	4.10	4.56	4.37	4.11
224	4.11	3.82	3.29	3.60	3.99	3.82	3.38
…	…	…	…	…	…	…	…
424	4.24	4.38	4.35	4.48	4.15	4.50	4.33

X_1 ~内容组织的合理性；X_2 ~问题展开的逻辑性；
X_3 ~回答学生的有效性；X_4 ~课下交流的有助性；
X_5 ~教材的帮助性；X_6 ~考试的公正性；Y ~总体评价.

建立Y与X_1~ X_6间简单、有效的模型，给教师提出建议.

问题分析



从X_1~ X_6中挑选出对Y影响显著的变量建立回归模型.

将所有对Y影响显著的X都选入模型，而影响不显著的X都不选入模型，使模型中自变量个数尽可能少.

解决办法

利用逐步回归

- 确定一个包含若干X的初始集合S_0 .
- 从S_0外的X中引入一个对Y影响最大的, $S_0 \to S_1$.
- 对S_1中的X进行检验，移出一个影响最小的, $S_1 \to S_2$.
- 继续进行，直到不能引入和移出为止 .
- 引入和移出都以给定的显著 性水平为标准.

MATLAB统计工具箱中的逐步回归

stepwise (x,y,inmodel,penter,premove)

输入x~自变量集合的$n \times k$ 数据矩阵（n是数据容量, k是变量数目）， y~因变量数据向量（n维）

Inmodel~初始模型S_0中包括的自变量集合的指标（即矩阵x的列序数，缺省时为无自变量）

penter~引入变量的显著性水平（缺省时为0.05）

premove~移出变量的显著性水平（缺省时为0.10）

输出几个交互式画面，供使用者人工选择变量，进行统计分析.

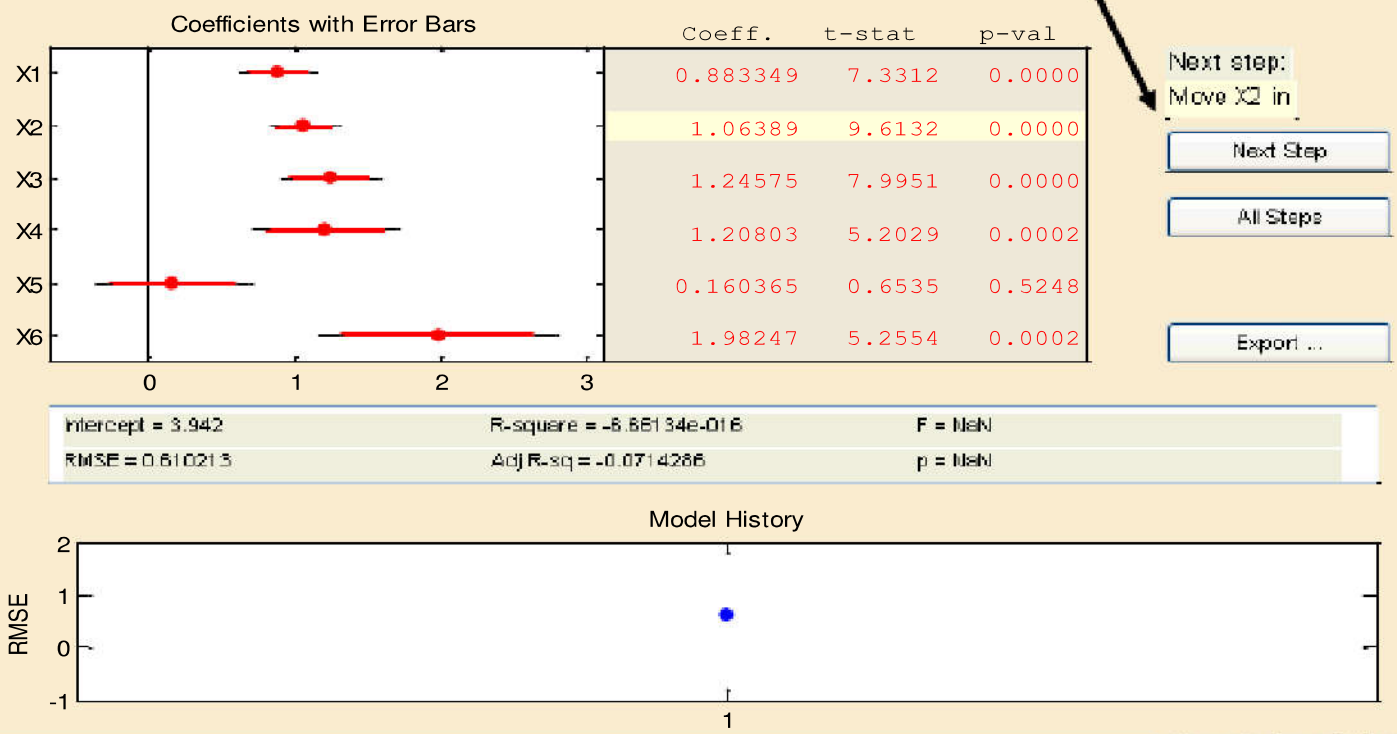
MATLAB统计工具箱中的逐步回归

stepwise (x,y)

其中x为X_1~ X_6数据矩阵, y为Y 向量

输出交互式画面

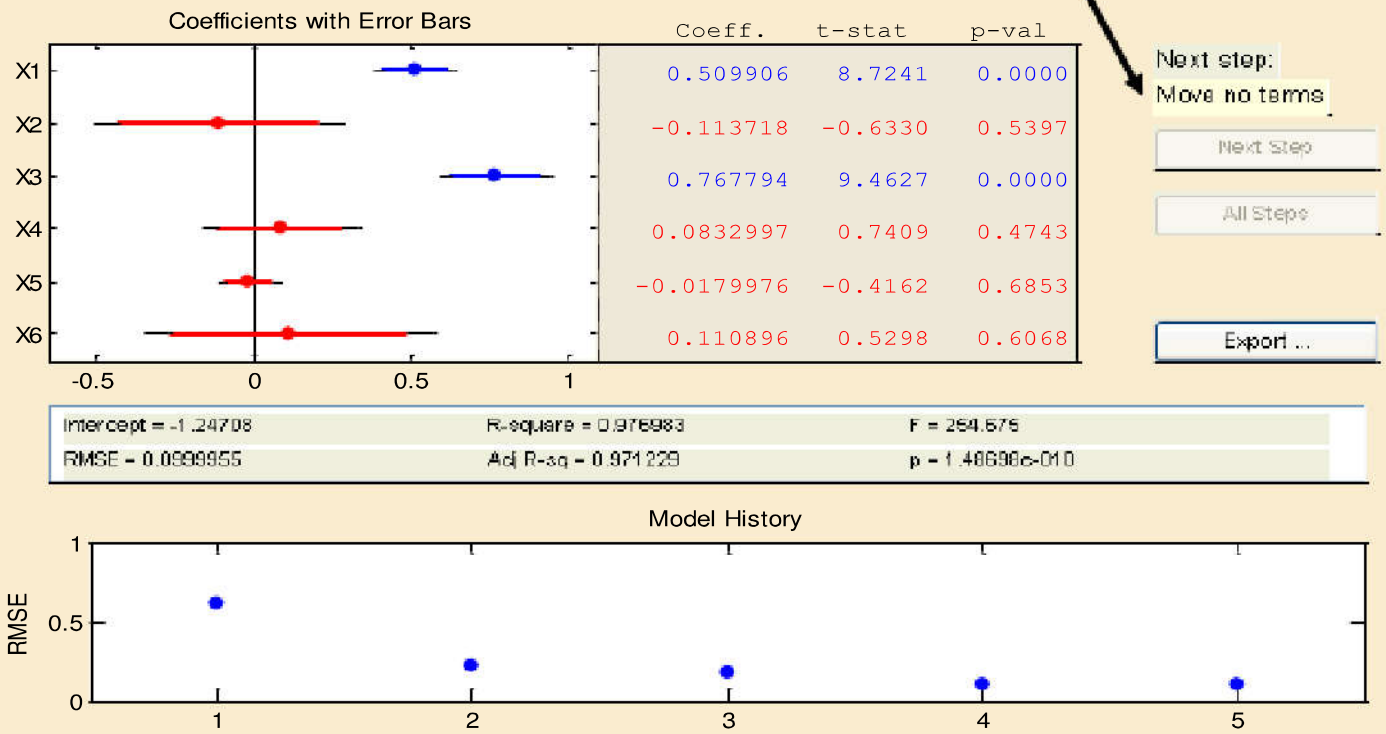
按照提示点击x2的热点引入x2



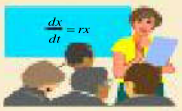
MATLAB统计工具箱中的逐步回归

依次按照提示: Move x3 in, Move x1 in, Move x2 out

按照提示, 包含x1,x3的模型是最终结果



最终模型 $Y = 0.5099\,X_1 + 0.7678\,X_3 - 1.2471$



模型解释 为什么只有X_1, X_2进入最终模型?

计算X_1~ X_6, Y 的相关系数矩阵(MATLAB的`corrcoef`):

1.0000	0.9008	0.6752	0.7361	0.2910	0.6471	0.8973
0.9008	1.0000	0.8504	0.7399	0.2775	0.8026	0.9363
0.6752	0.8504	1.0000	0.7499	0.0808	0.8490	0.9116
0.7361	0.7399	0.7499	1.0000	0.4370	0.7041	0.8219
0.2910	0.2775	0.0808	0.4370	1.0000	0.1872	0.1783
0.6471	0.8026	0.8490	0.7041	0.1872	1.0000	0.8246
0.8973	0.9363	0.9116	0.8219	0.1783	0.8246	1.0000

- 与Y的相关系数大于0.85的是X_1, X_2, X_3 .
- X_2与X_1, X_3 的相关系数大于0.85.

模型解释

$$Y = 0.5099\,X_1 + 0.7678\,X_3 - 1.2471$$

X_1 ~内容组织的合理性；X_2 ~问题展开的逻辑性；

X_3 ~回答学生的有效性；X_4 ~课下交流的有助性；

X_5 ~教材的帮助性；X_6 ~考试的公正性；Y ~总体评价.

X_1提高1分Y提高0.5分，X_3提高1分Y提高0.77分.

逐步回归

- 逐步回归是从众多变量中挑选出影响显著变量的有效方法.
- 原有变量的平方项、交互项等也可以作为新变量加入到候选行列，用逐步回归处理.

10.6 冠心病与年龄

- 冠心病是一种常见的心脏疾病，严重危害人类的健康.
- 多项研究表明，冠心病发病率随着年龄的增加而上升.
- 在冠心病流行病学研究中年龄是最常见的混杂因素之一.

100名被观察者的年龄及他们是否患冠心病的数据

序号	年龄	冠心病	序号	年龄	冠心病	序号	年龄	冠心病	序号	年龄	冠心病
1	20	0	26	35	0	51	44	1	76	55	1
…	…	…	…	…	…	…	…	…	…	…	…
25	34	0	50	44	0	75	55	1	100	69	1

根据以上数据建立数学模型，分析发病率与年龄的关系，并进行统计预测.

分析与假设

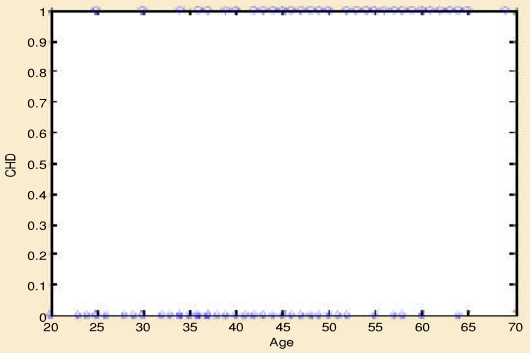
被观察者独立选取

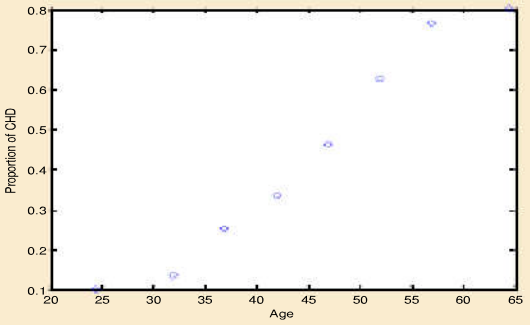
x~被观察者年龄, Y~患病情况 (Y=1~患病, Y=0~不患病)

无法建立前面那样的回归模型,
需要对数据进行预处理.

按年龄段分组统计患病人数及比例

年龄段	段中点	人数	患病人数	患病比例
20-29	24.5	10	1	0.1
30-34	32	15	2	0.13
…	…	…	…	…
60-69	64.5	10	8	0.80
合计		100	43	0.43



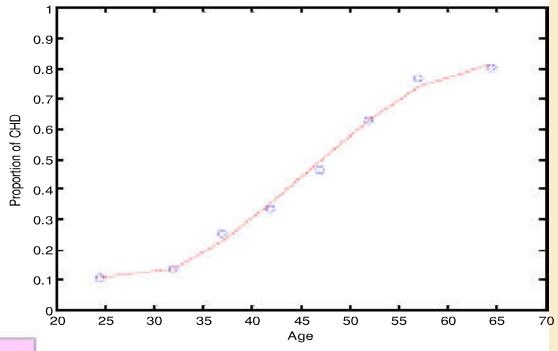


患病比例随年龄增大而递增,是介于0与1之间的S型曲线.

分析与假设

Y的条件期望 $y = E(Y \mid x)$　　**Y取值 0, 1 ; y 取值 [0, 1]**

用普通方法建立回归方程　$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$



- y取值不一定在[0,1]中.
- 误差项ε只能取值0,1, 不具有正态性, 且具有异方差性.

违反普通回归分析的前提条件!

当因变量Y为一个二分类(或多分类)变量时，需要用到新的回归模型.

Logit 模型

$\pi(x)$~年龄x的患病概率(患病比例y)　$\pi(x)=P(Y=1\mid x)$

Y的(条件)期望 $y=E(Y\mid x)$　　方差 $D(Y\mid x)=\pi(x)(1-\pi(x))$

$\pi(x)$ ~ S型曲线, 取值[0,1]　⇨ **Logistic模型**

$$\pi(x)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

⇩ 反函数

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)=\beta_0+\beta_1 x$$

$\pi(x)$的变换 $\mathrm{Logit}(\pi(x))=\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$

连接函数，
取值 $(-\infty,+\infty)$

Logit模型 (Logistic回归模型)

Logit 模型

数据预处理: 将年龄分成k(=8)组.

x_i~第i组年龄, n_i~被观察人数, m_i~患病人数, i=1,…, k

患病概率 $\pi_i = m_i / n_i$

Logit 模型 $\mathrm{Logit}(\pi_i) = \ln(\dfrac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_i$

β_0,β_1~回归系数　　设m_i服从二项分布 $B(n_i, \pi_i)$

回归系数可用极大似然法估计得到.

模型求解

Logit模型可用MATLAB命令glmfit求解

b = glmfit(x, y, 'distr', 'link')

[b,dev,stats] = glmfit(x, y, 'distr', 'link')

x~自变量数据矩阵(第1列自动添加列向量1).

y~因变量数据向量(对distr =binomial, y可取矩阵:第1列为“成功”次数, 第2列为观察次数).

'distr' ~估计系数所用分布('binomial','poisson' 等), 缺省时为 ' normal' .

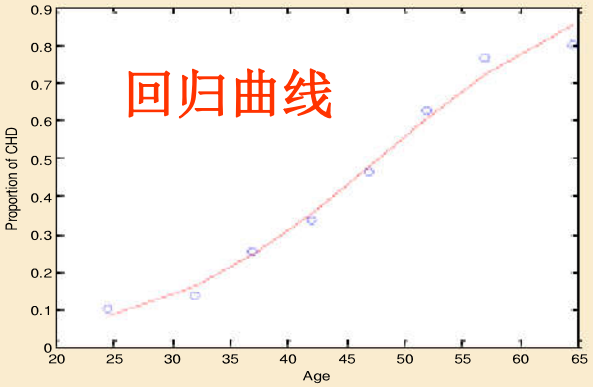
'link' ~'logit','probit' 等(缺省时为' logit').

b~回归系数的估计值, dev~拟合偏差, stats~统计指标

模型求解

编程计算

参数	估计值	标准差
β_0	-5.0382	1.0863
β_1	0.1050	0.0231



拟合偏差0.5242

[yhat, dylo, dyhi] = glmval(b, x, 'logit')

自变量为x时y的预测值yhat及置信度为95%的置信区间

年龄段	年龄x	患病比例（实际值）	患病比例（预测值y）	置信区间
20-29	24.5	0.1	0.0783	[0.0282, 0.1992]
…	…	…	…	…
60-69	64.5	0.80	0.8501	[0.6855, 0.9366]

模型评价与结果分析

- **Logit模型是否需要引入x^2项？**

$$\text{Logit}(\pi(x)) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \quad \Rightarrow \quad \beta_0 + \beta_1 x + \beta_2 x^2$$

用似然比统计量计算： **pval = 1 - chi2cdf(dev-dev2,1) =0.9371**

⇒ 模型中引入x^2项不能显著提高拟合程度.

- 选用**Probit模型**(另一种广义线性模型)结果如何？

$$\pi(x) = \Phi(\beta_0 + \beta_1 x) \qquad \text{Probit}(\pi(x)) = \Phi^{-1}(\pi(x)) = \beta_0 + \beta_1 x$$

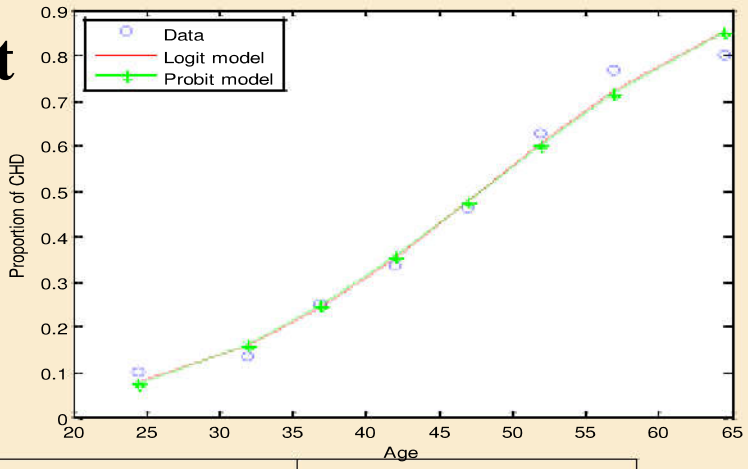
Φ是正态概率分布函数(S型曲线)

模型评价与结果分析

Probit模型求解

glmfit中需将logit改为probit

参数	估计值	标准差
β_0	-2.9933	0.6011
β_1	0.0624	0.0128



拟合偏差0.6529

年龄段	年龄x	患病比例（实际值）	预测值1（Logit）	预测值2（Probit）
20-29	24.5	0.1	0.0783	0.0715
…	…	…	…	…
60-69	64.5	0.80	0.8501	0.8489

两个模型的拟合程度不相上下.

模型评价与结果分析

β_1的直观解释

$$\text{Logit}(\hat{\pi}(x)) = \ln\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x = -5.0382 + 0.1050\,x$$

Odds~事件发生(患病)概率与不发生(不患病)概率之比.

$$\text{Odds}(x) = \frac{\pi(x)}{1-\pi(x)} = e^{(\beta_0+\beta_1 x)}$$ 年龄x的人患病与不患病概率之比

年龄增加1岁的Odds比(发生比率)

$$\frac{\text{Odds}(x+1)}{\text{Odds}(x)} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1} \Rightarrow \beta_1 = \ln\left(\frac{\text{Odds}(x+1)}{\text{Odds}(x)}\right)$$

年龄增加1岁Odds比的对数

$$\text{Odds}(x+k) = e^{k\beta_1}\text{Odds}(x)$$

年龄增加k岁后的Odds

模型评价与结果分析

$$\text{Logit}(\hat{\pi}(x)) = \ln\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x = -5.0382 + 0.1050\,x$$

20岁的青年人患冠心病的概率 $\hat{\pi}(20) = 0.0503$

发生比(患与不患冠心病的概率之比） $\text{Odds}(20) = 0.0593$

⇨ 年龄增加1岁患病概率的变化很小.

10年后30岁人的发生比 $\text{Odds}(30) = e^{10\times\beta_1} \times 0.0593 = 0.1694$

60岁时 $\text{Odds}(60) = 3.9545$ 是20岁的 $e^{40\times\beta_1} = 66.6863$ 倍

Logit回归模型 $\pi(x) = 0.5 \Rightarrow x = ?$ $\hat{\pi}(x^*) = 0.5$ ⇨ $x^* = 48$

⇨ 48岁时患冠心病的概率会大于不患冠心病的概率.

模型评述

• 因变量是定性变量的回归分析作为一种有效的数据处理方法已被广泛应用，尤其在医学、社会调查、生物信息处理等领域 .

多元Logit模型 $\text{Logit}(\pi(x)) = \ln(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$

自变量 $x_1, \cdots, x_m$ 可以是定量变量或定性变量

• 可以用逐步回归方法建立多元Logit模型和Probit模型, 逐个地加入自变量(包括某个自变量的高次项及某些自变量的交叉变量), 并且实时地进行模型比较检验, 选择与数据拟合较好的模型.