中国科学技术大学
University of Science and Technology of China

数学建模
Mathematical Modeling

陈仁杰
中国科学技术大学

统计回归模型

回归分析

- 研究因变量对自变量的依赖关系的一种统计分析方法
- 通过自变量的给定值来估计或预测因变量的均值
- 可用于预测、时间序列建模以及发现各种变量之间的因果关系
- 本质：数据拟合（函数拟合）+ 统计分析（显著性检验、区间估计等）

回归分析的作用

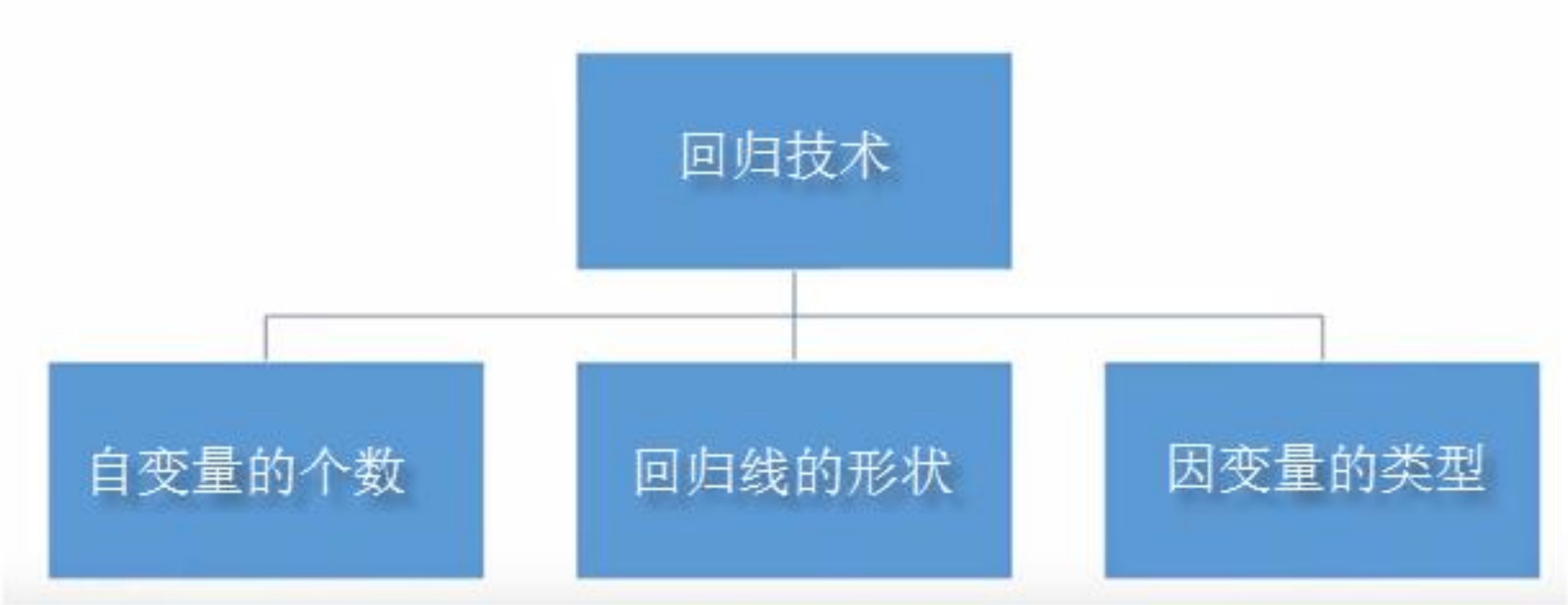
1）挑选与因变量相关的自变量;

2）描述因变量与自变量之间的关系强度;

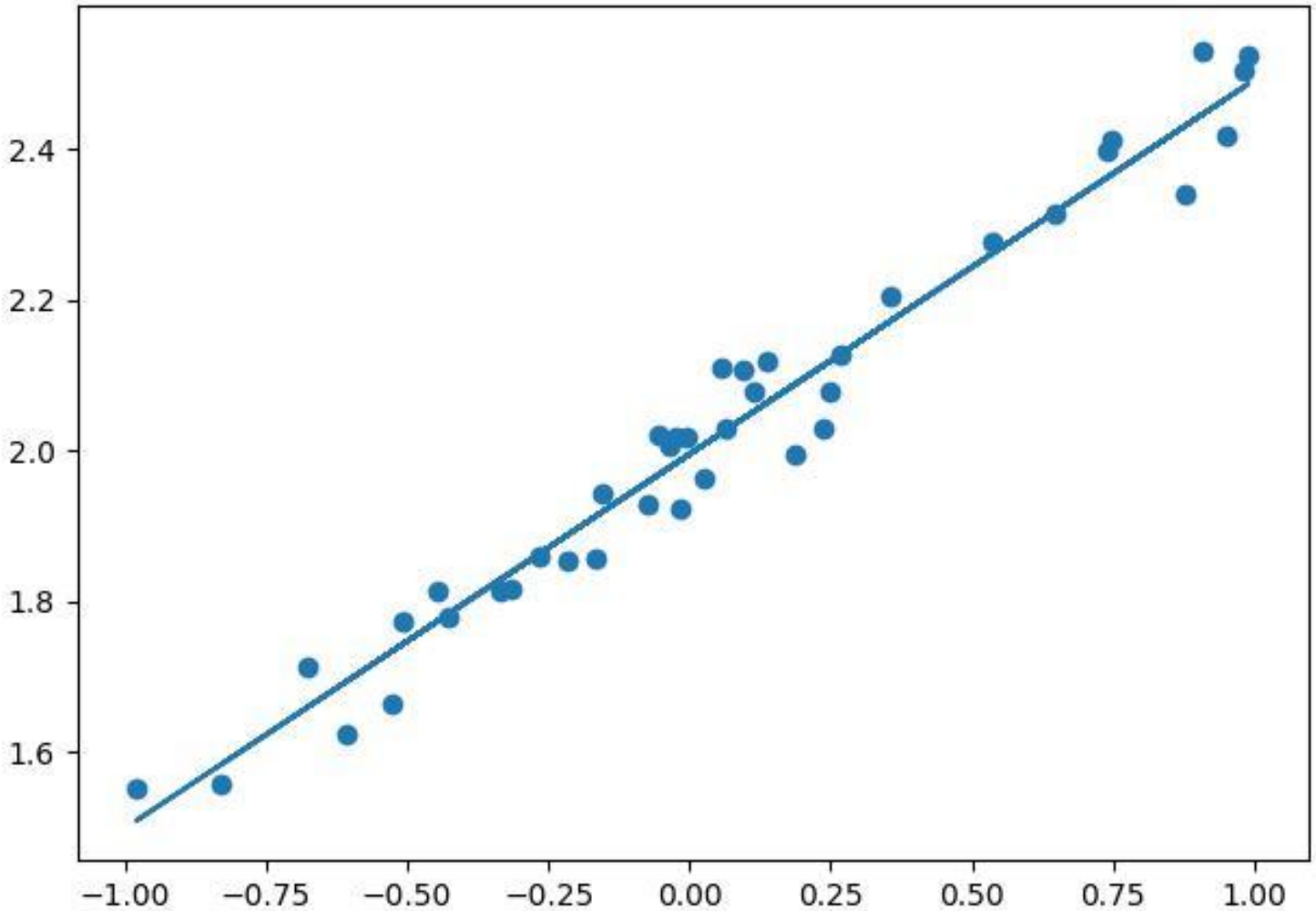
3）生成模型，通过自变量来预测因变量;

4）根据模型，通过因变量，来控制自变量。

回归分析方法



1. 线性回归



线性回归

1）自变量与因变量之间必须有线性关系；

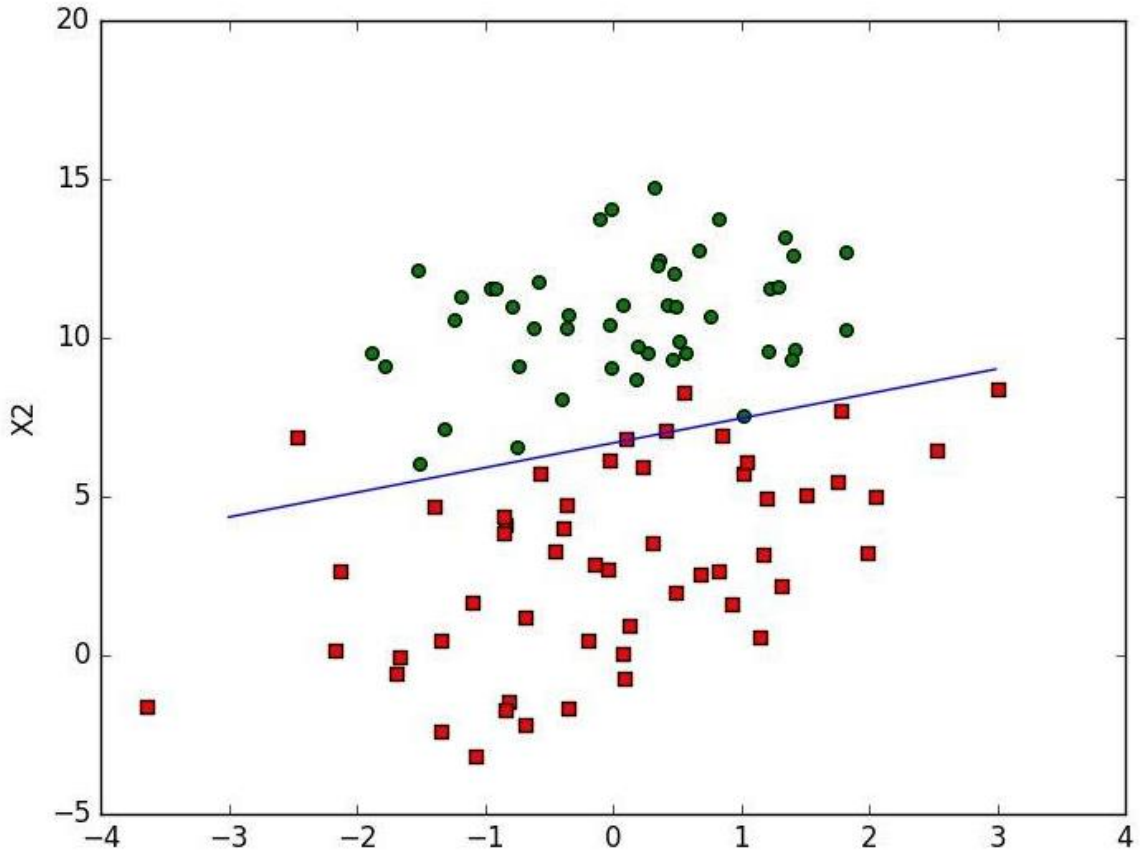
2）多元回归存在多重共线性，自相关性和异方差性；

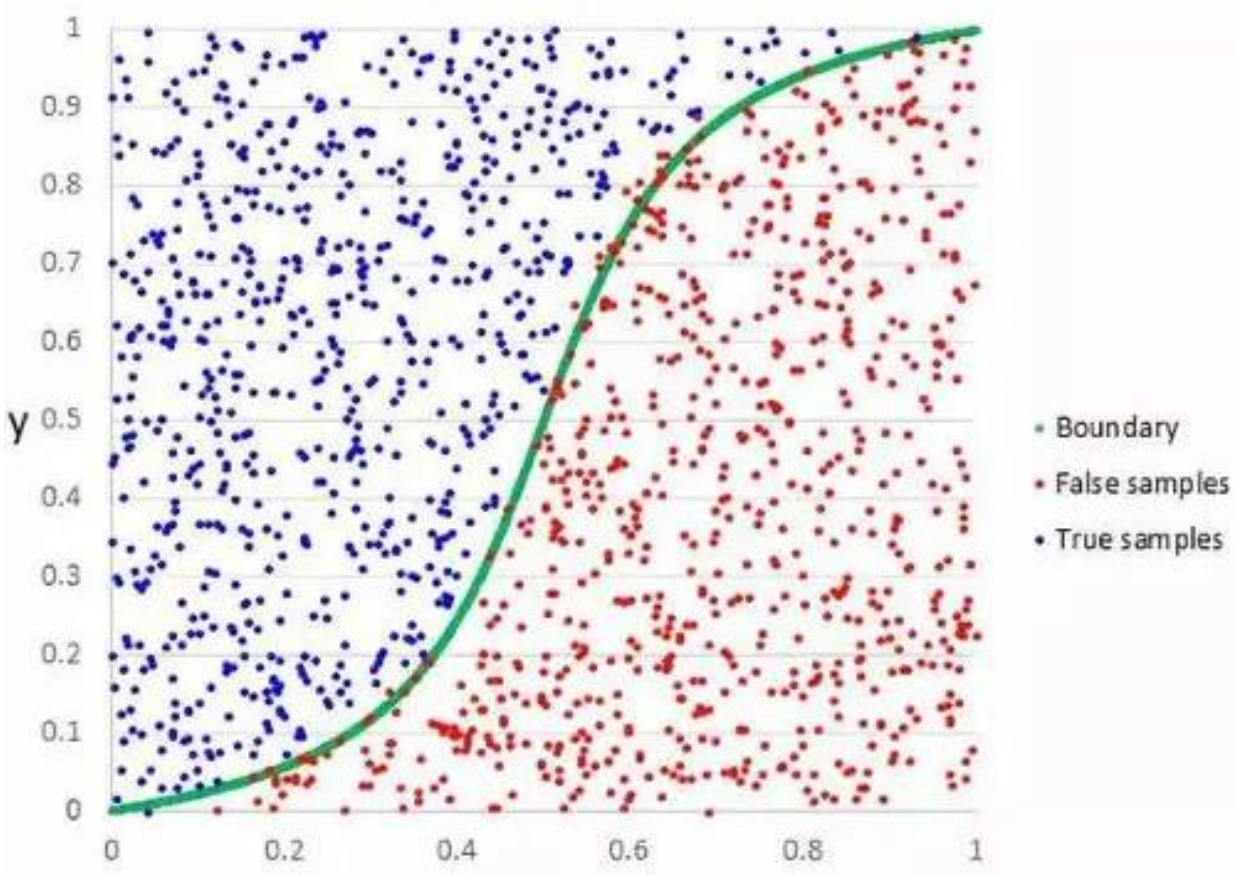
3）线性回归对异常值非常敏感。它会严重影响回归线，最终影响预测值；

4）多重共线性会增加系数估计值的方差，使得估计值对于模型的轻微变化异常敏感，结果就是系数估计值不稳定；

5）在存在多个自变量的情况下，可以使用向前选择法，向后剔除法和逐步筛选法来选择最重要的自变量。

2. Logistic回归





Logistic回归

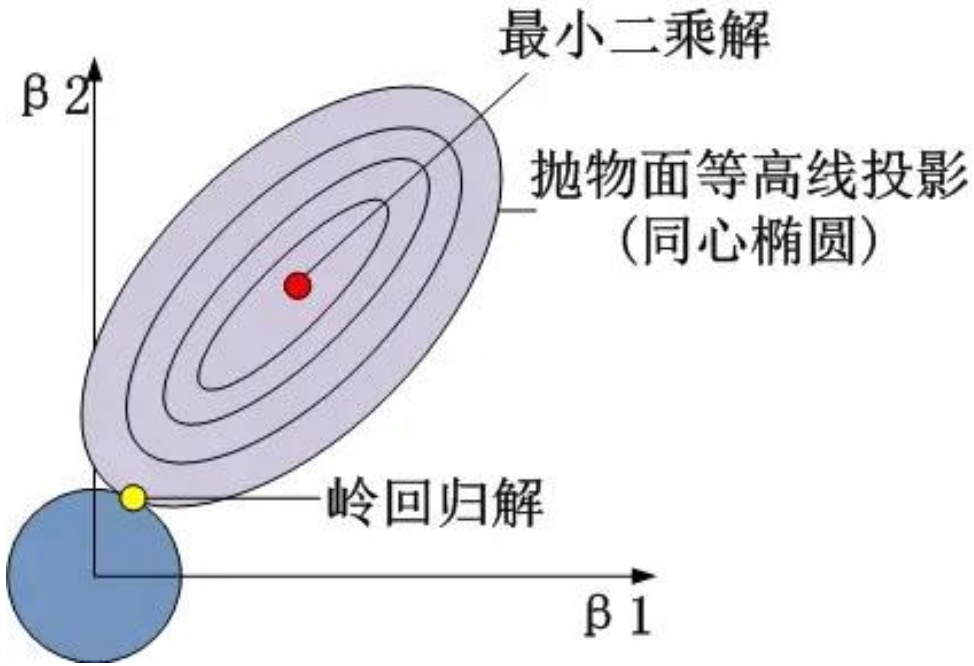
- Logistic回归广泛用于分类问题;
- Logistic回归不要求自变量和因变量存在线性关系。它可以处理多种类型的关系，因为它对预测的相对风险指数使用了一个非线性的 log 转换;
- Logistic回归需要较大的样本量，因为在样本数量较少的情况下，极大似然估计的效果比普通最小二乘法差;
- 自变量之间应该互不相关，即不存在多重共线性。然而，在分析和建模中，我们可以选择包含分类变量相互作用的影响;
- 如果因变量的值是定序变量，则称它为序Logistic回归.

3. 岭回归 (ridge regression)

- 又称脊回归、吉洪诺夫正则化（Tikhonov regularization）
- 目的：给回归估计值添加一个偏差值，来降低标准误差（方差值）

$$= \arg\min \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

$$= \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$



岭回归

1）除常数项以外，岭回归的假设与最小二乘回归相同；

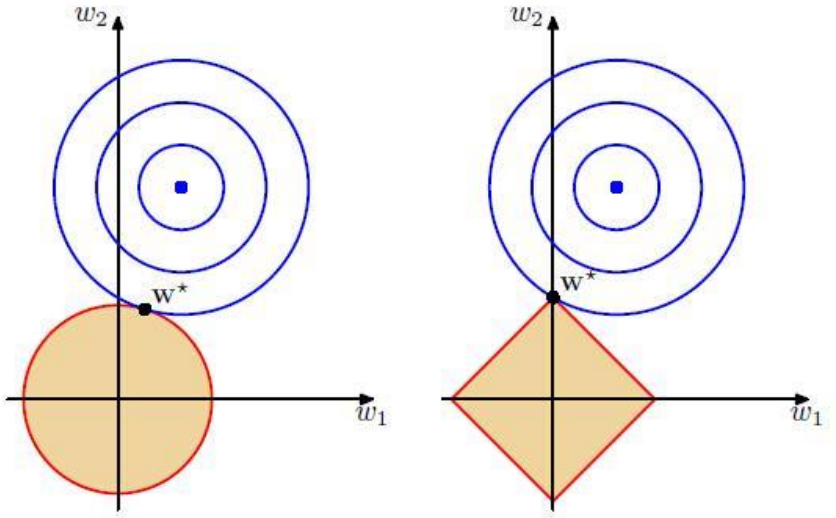
2）它收缩了相关系数的值，但没有达到零，这表明它不具有特征选择功能；

3）这是一个正则化方法，并且使用的是 L2 正则化。

4.套索回归 (Lasso regression)

- 目标：使一些参数估计结果等于零。使用的惩罚值越大，估计值会越趋近于零。
- 从给定的n个变量之中选择变量（特征选择）

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$



套索回归

1）除常数项以外，这种回归的假设与最小二乘回归类似;

2）它将收缩系数缩减至零（等于零），这确实有助于特征选择;

3）这是一个正则化方法，使用的是 L1 正则化;

4）如果一组预测因子是高度相关的，套索回归会选出其中一个因子并且将其它因子收缩为零。

相关模型

- 稀疏优化
- 压缩感知

5. ElasticNet 回归

ElasticNet 回归是套索回归和岭回归的组合体。它会事先使用L1和L2作为正则化矩阵进行训练

1）在高度相关变量的情况下，它会产生群体效应；

2）选择变量的数目没有限制；

3）它可以承受双重收缩。

如何选择回归模型?

- 数据探索：识别变量的关系和影响
- 比较不同模型的拟合优点：分析不同的指标参数，如统计意义的参数，R-square，调整 R-square，AIC，BIC以及误差项，另一个是 Mallows’ Cp 准则
- 交叉验证：评估预测模型最好的方法。需将数据集分成两份（一份用于训练，一份用于验证）。使用观测值和预测值之间的均方差即可快速衡量预测精度。
- 回归正则化方法（套索，岭和ElasticNet）在高维数据和数据集变量之间存在多重共线性的情况下运行良好。

诊断回归分析结果

1.自变量与因变量是否具有预期的关系

2.自变量对模型是否有帮助

3.残差是否有空间聚类

4.模型是否出现了倾向性

5.自变量中是否存在冗余

6.评估模型的性能

https://www.sohu.com/a/165006005_197042

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

统计回归模型

1　牙膏的销售量

2　软件开发人员的薪金

3　酶促反应

4　投资额与国民生产总值和物价指数

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，及同期其它厂家同类牙膏的平均售价

销售周期	本公司价格(元)	其它厂家价格(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

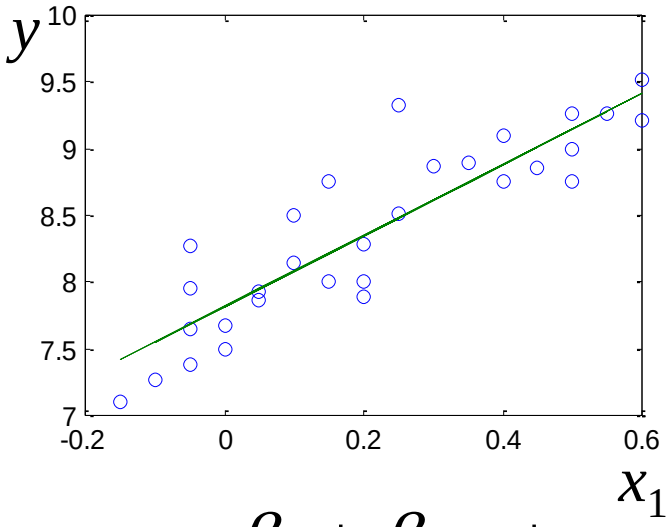
基本模型

y ~公司牙膏销售量

x_1~其它厂家与本公司价格差

x_2~公司广告费用

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



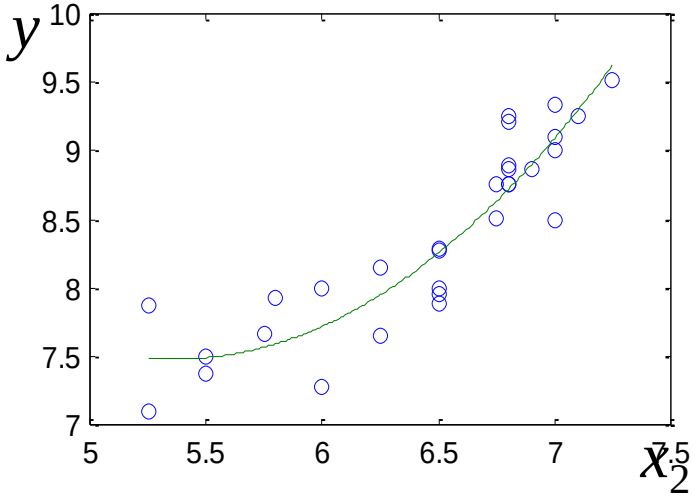
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

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=0.0000		

Stats~检验统计量

R^2, F, p

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=0.0000		

y的90.54%可由模型确定

F远超过F检验的临界值

p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

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

价格差x_1=其它厂家价格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=650万元

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 = 8.2933 \text{ (百万支)}$$

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 7.8320×3.7≈ 29（百万元）以上

模型改进

x_1和x_2对y的影响独立

x_1和x_2对y的影响有交互作用

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

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=0.0000		

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

参数	参数估计值	置信区间
β_0	29.1133	[13.7013 44.5252]
β_1	11.1342	[1.9778 20.2906]
β_2	-7.6080	[-12.6932 -2.5228]
β_3	0.6712	[0.2538 1.0887]
β_4	-1.4777	[-2.8518 -0.1037]
R^2=0.9209 F=72.7771 p=0.0000		

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

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

$\hat{y}=8.2933$ (百万支)

区间 [7.8230，8.7636]

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

$\hat{y}=8.3272$ (百万支)

区间 [7.8953，8.7592]

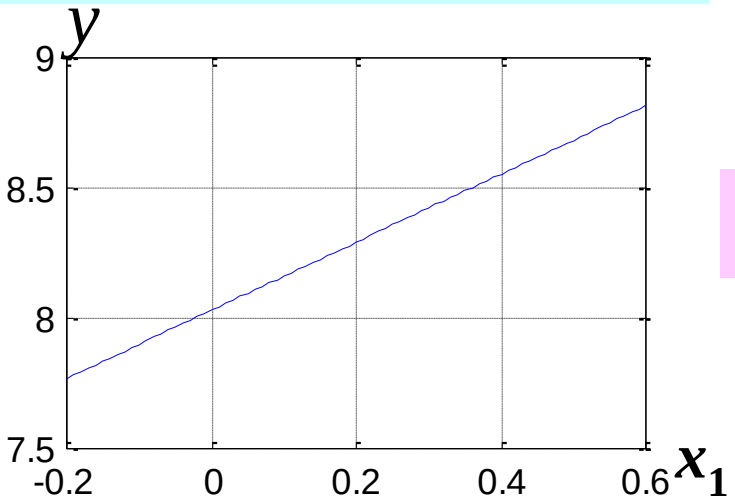
$\hat{y}$ 略有增加

预测区间长度更短

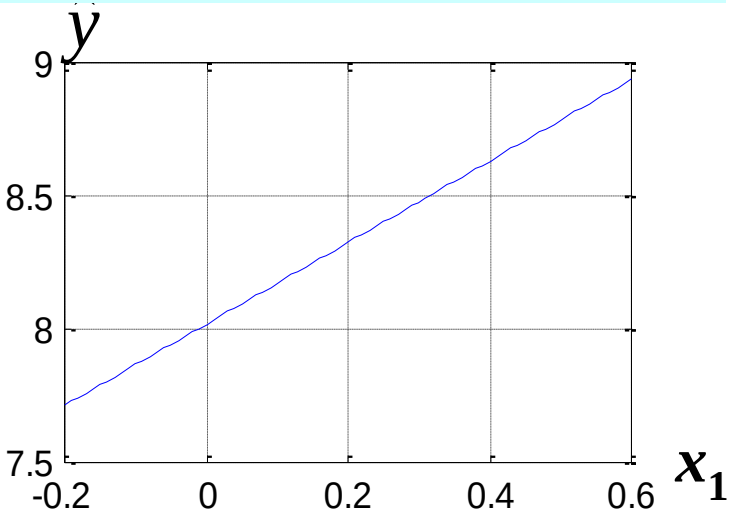
两模型$\hat{y}$ 与x_1,x_2关系的比较

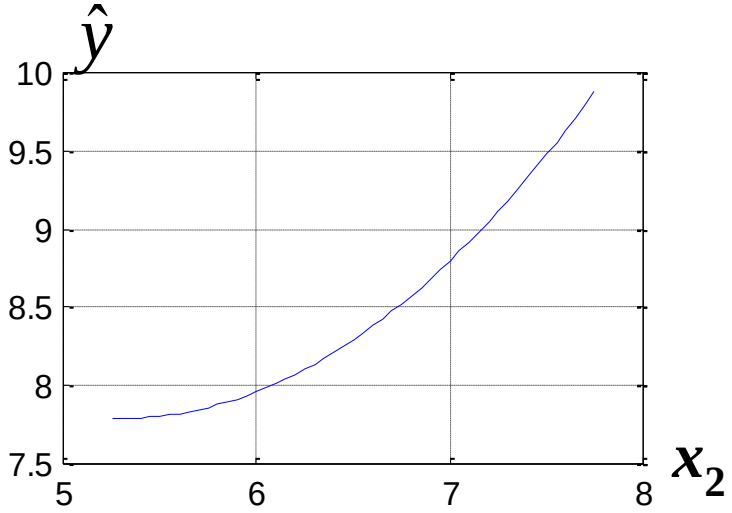
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2$$

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

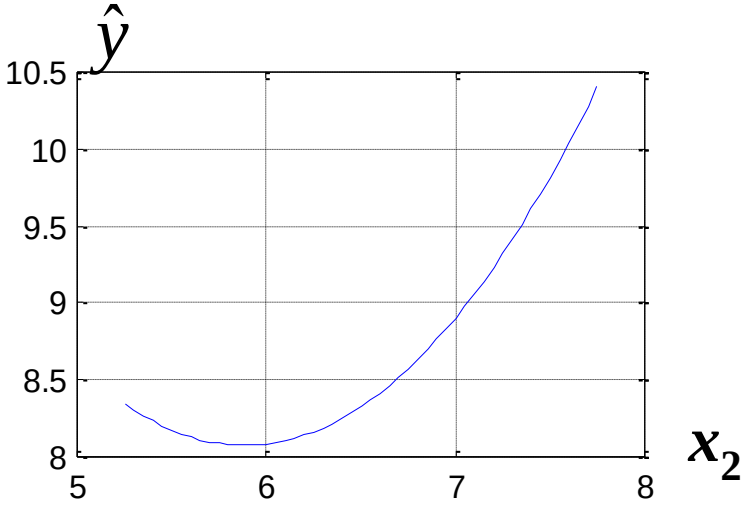


x_2=6.5





x_1=0.2



交互作用影响的讨论

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

价格差 x_1=0.1

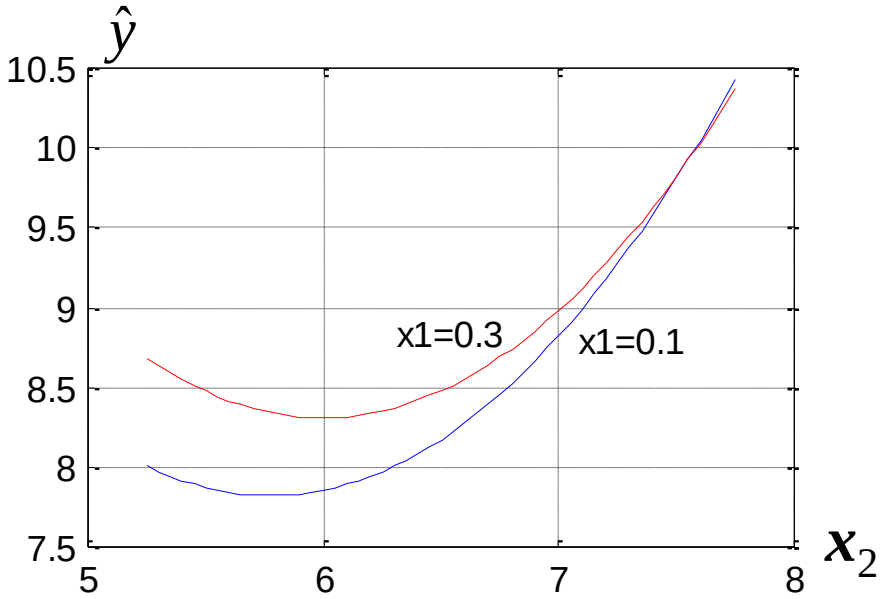
$$\hat{y}\big|_{x_1=0.1} = 30.2267 - 7.7558x_2 + 0.6712x_2^2$$

价格差 x_1=0.3

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

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

⇨ **价格优势会使销售量增加**



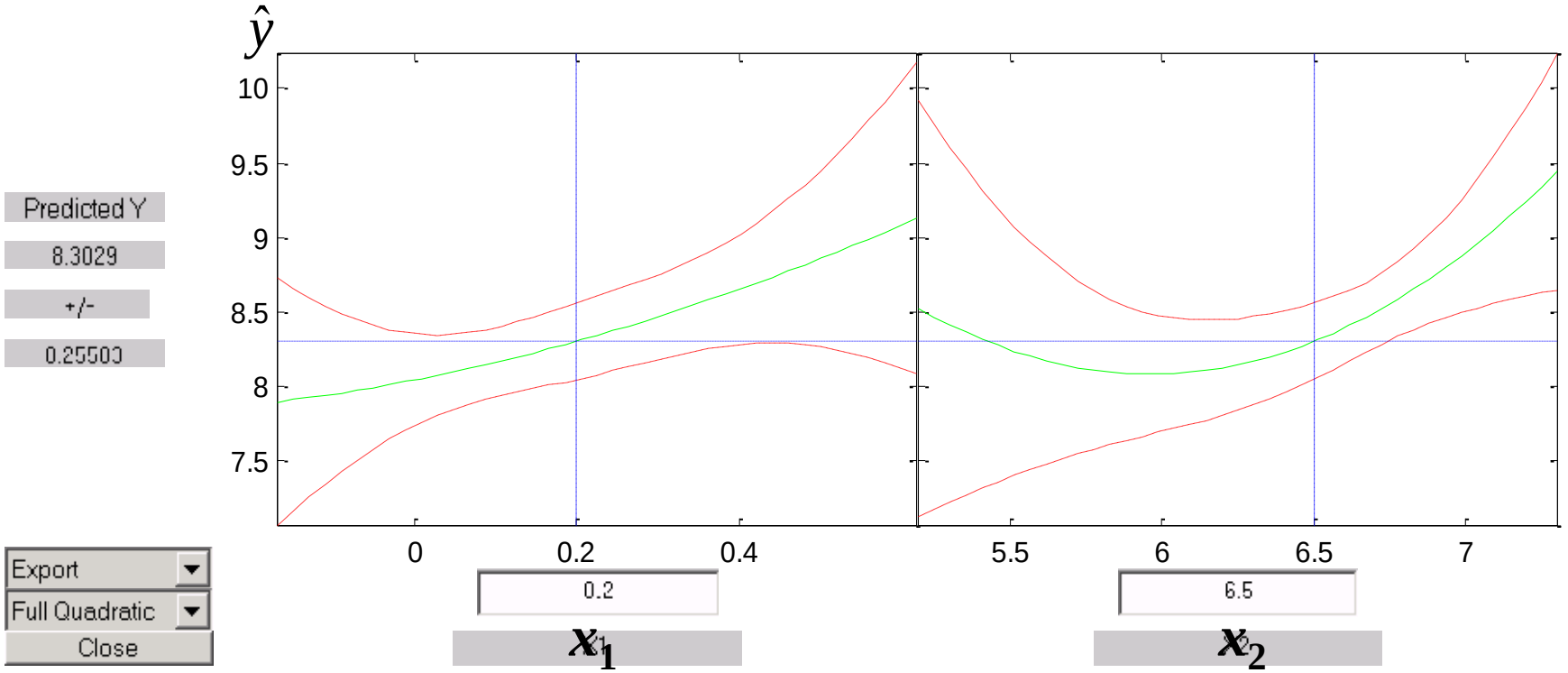
加大广告投入使销售量增加
（x_2大于6百万元）

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

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

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

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

⇨

中学：$x_3=1, x_4=0$ ；

大学：$x_3=0, x_4=1$；

更高：$x_3=0, x_4=0$

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

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + \varepsilon$$

参数	参数估计值	置信区间
a_0	11032	[10258 11807]
a_1	546	[484 608]
a_2	6883	[6248 7517]
a_3	-2994	[-3826 -2162]
a_4	148	[-636 931]
R^2=0.957	F=226	p=0.000

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

$R^2, F, p \rightarrow$ 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：$x_3=0, x_4=0$.

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

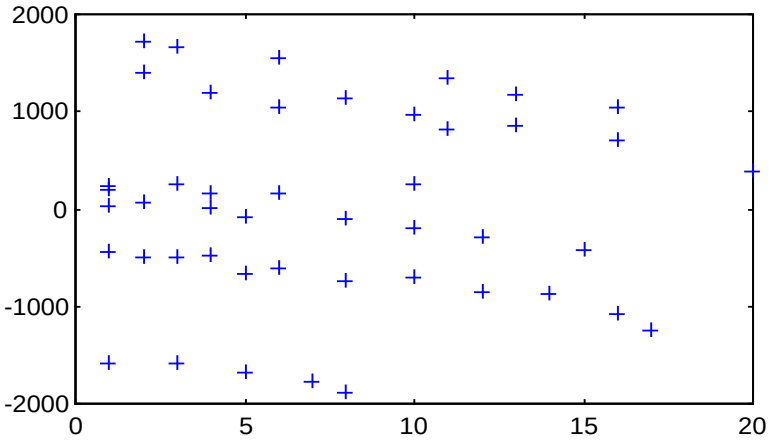
$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \hat{a}_3 x_3 + \hat{a}_4 x_4$$

残差 $e = y - \hat{y}$

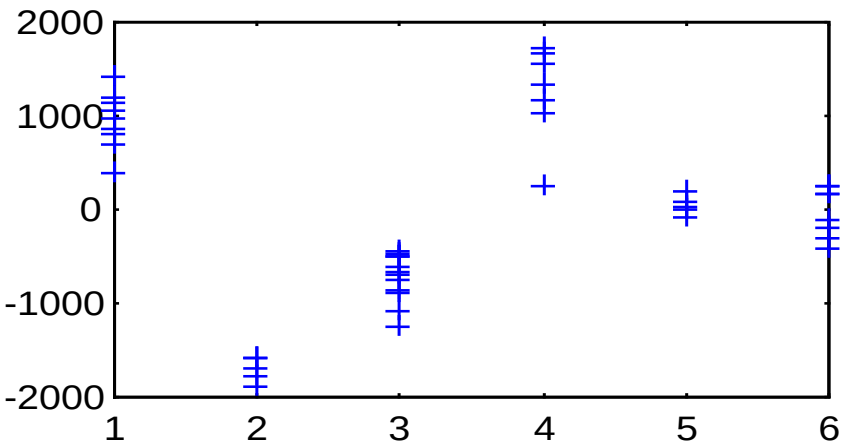
管理与教育的组合

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

e 与资历x_1的关系



e与管理—教育组合的关系



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 。

残差全为正，或全为负，管理—教育组合处理不当

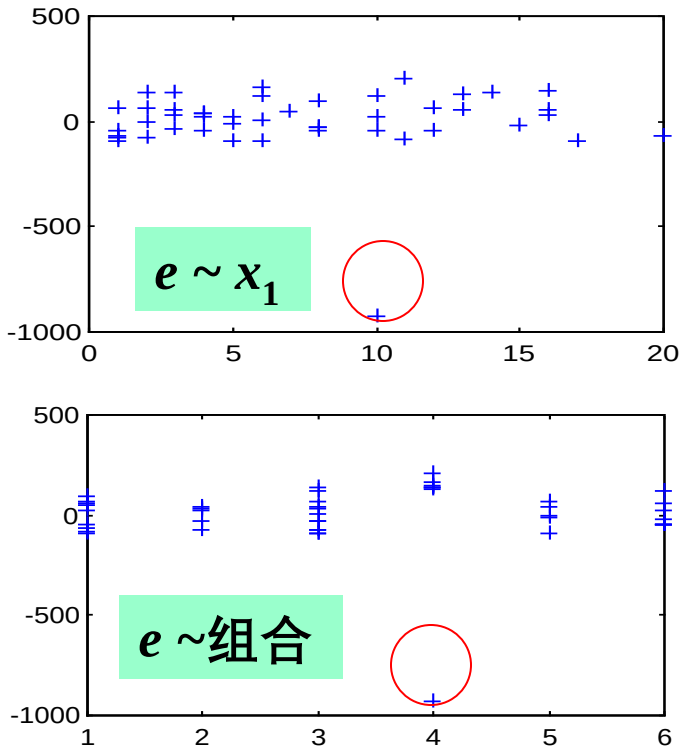
应在模型中增加管理x_2与教育x_3, x_4的交互项

进一步模型

增加管理x_2与教育x_3, x_4的交互项

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

参数	参数估计值	置信区间
a_0	11204	[11044 11363]
a_1	497	[486 508]
a_2	7048	[6841 7255]
a_3	-1727	[-1939 -1514]
a_4	-348	[-545 –152]
a_5	-3071	[-3372 -2769]
a_6	1836	[1571 2101]
R^2=0.999	F=554	p=0.000



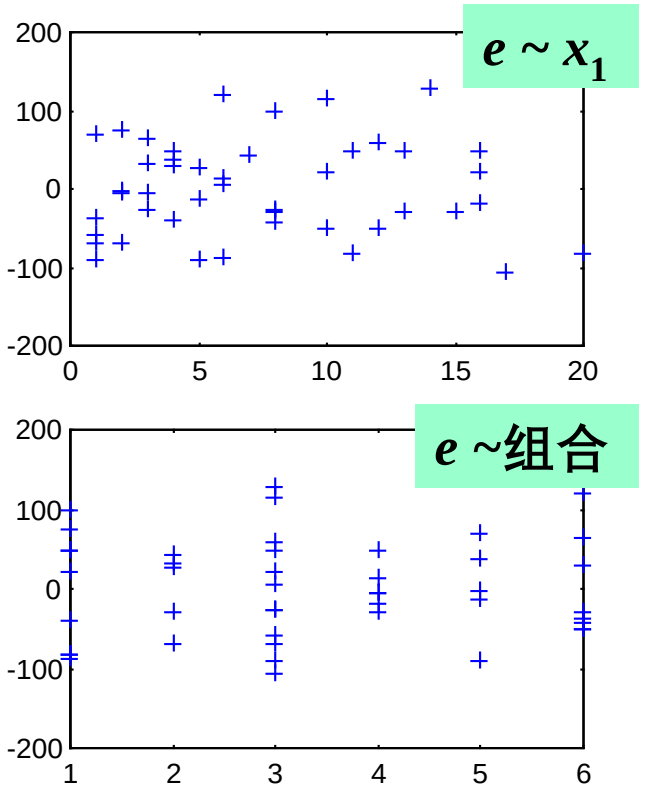
R^2, F有改进，所有回归系数置信区间都不含零点，模型完全可用

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

参数	参数估计值	置信区间
a_0	11200	[11139 11261]
a_1	498	[494 503]
a_2	7041	[6962 7120]
a_3	-1737	[-1818 -1656]
a_4	-356	[-431 –281]
a_5	-3056	[-3171 –2942]
a_6	1997	[1894 2100]
R^2= 0.9998 F=36701 p=0.0000		



R^2： 0.957 → 0.999 → 0.9998

F： 226 → 554 → 36701

置信区间长度更短

残差图十分正常

最终模型的结果可以应用

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变量的个数应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

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	/

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

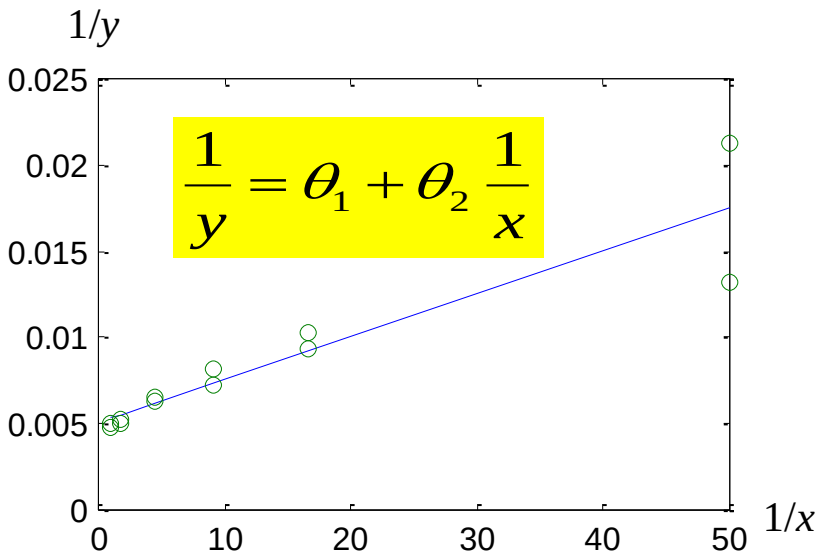
经噻吟霉素处理后实验数据的估计结果

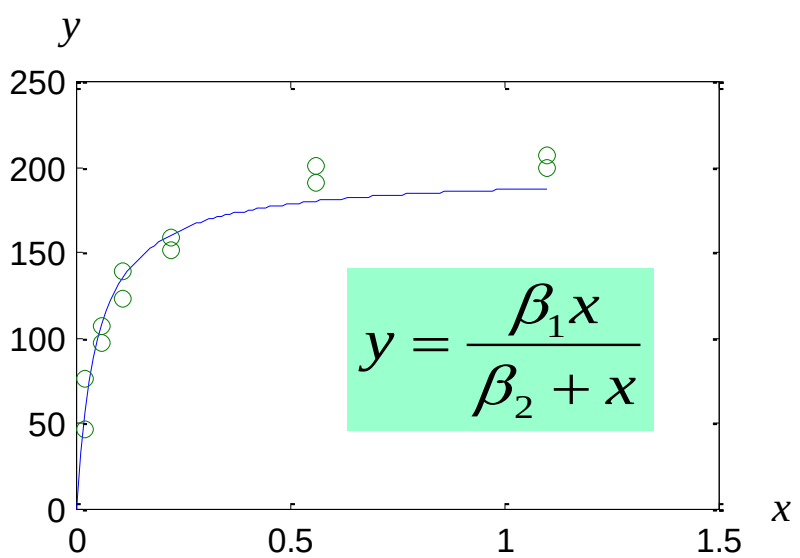
参数	参数估计值（×10^{-3}）	置信区间（×10^{-3}）
θ_1	5.107	[3.539 6.676]
θ_2	0.247	[0.176 0.319]
R^2=0.8557	F=59.2975	p=0.0000

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

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

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)

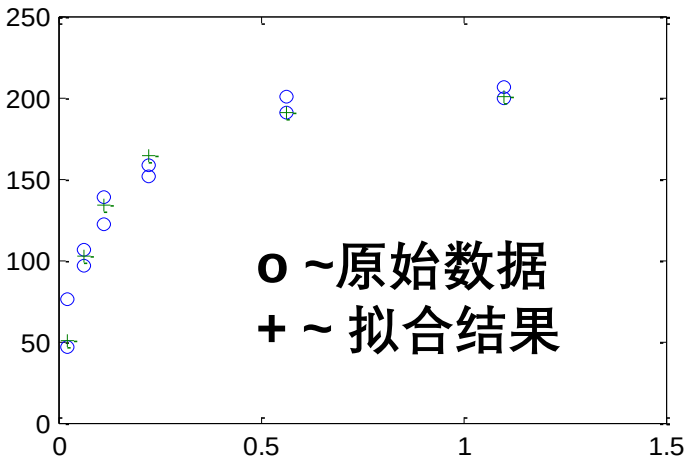
```
x=                    ; y=                    ;
beta0=[195.8027  0.04841];
[beta,R,J]=nlinfit(x,y,'f1',beta0);
betaci=nlparci(beta,R,J);
beta, betaci
```

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

非线性模型结果分析

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

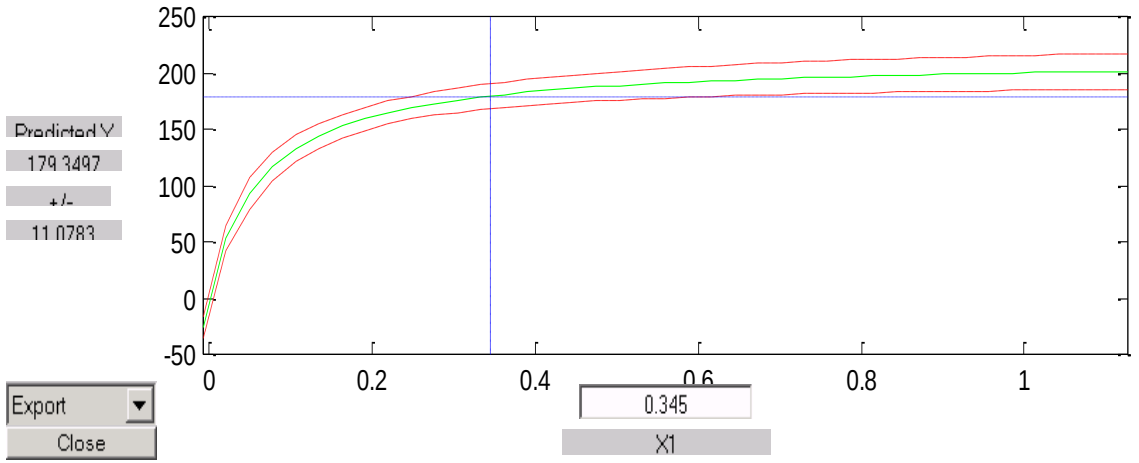
参数	参数估计值	置信区间
β_1	212.6819	[197.2029 228.1609]
β_2	0.0641	[0.0457 0.0826]



最终反应速度为 $\hat{\beta}_1 = 212.6831$

半速度点(达到最终速度一半时的x值)为

$$\hat{\beta}_2 = 0.0641$$



拖动画面的十字线，得y的预测值和预测区间

画面左下方的Export 输出其它统计结果。

剩余标准差s= 10.9337

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

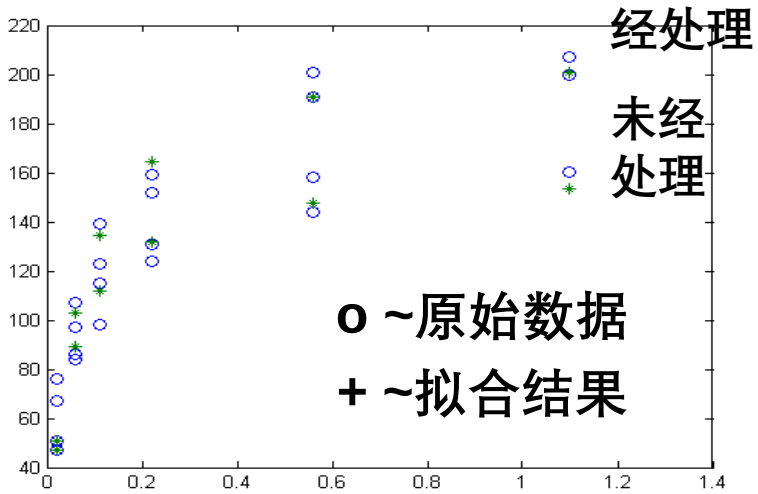
混合模型求解

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

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

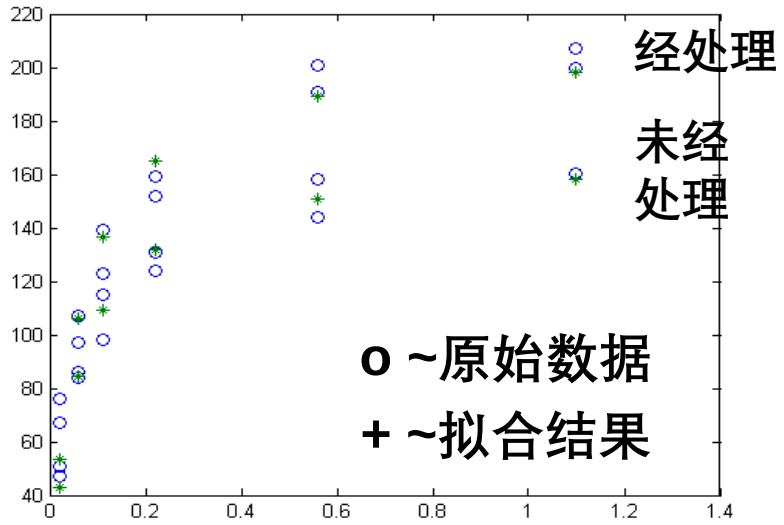
⇨ 经嘌呤霉素处理的作用不影响半速度点参数

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。

4 投资额与国民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

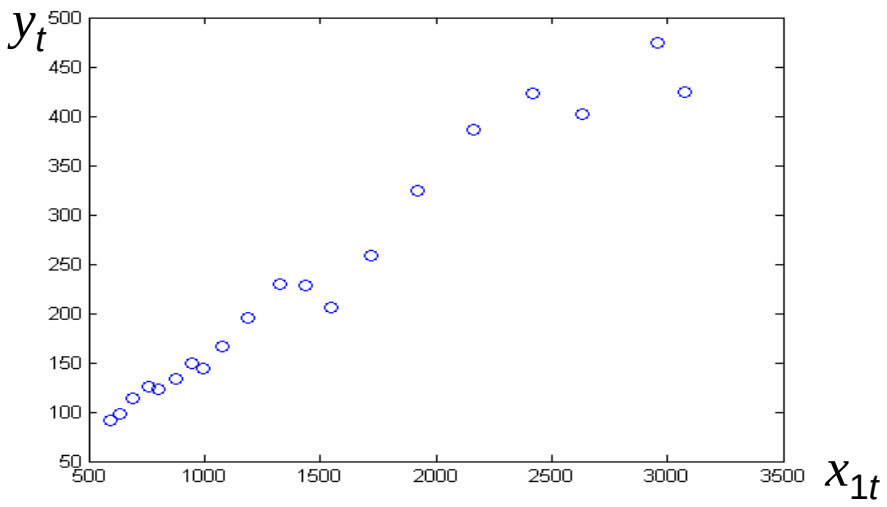
若采用普通回归模型直接处理，将会出现不良后果

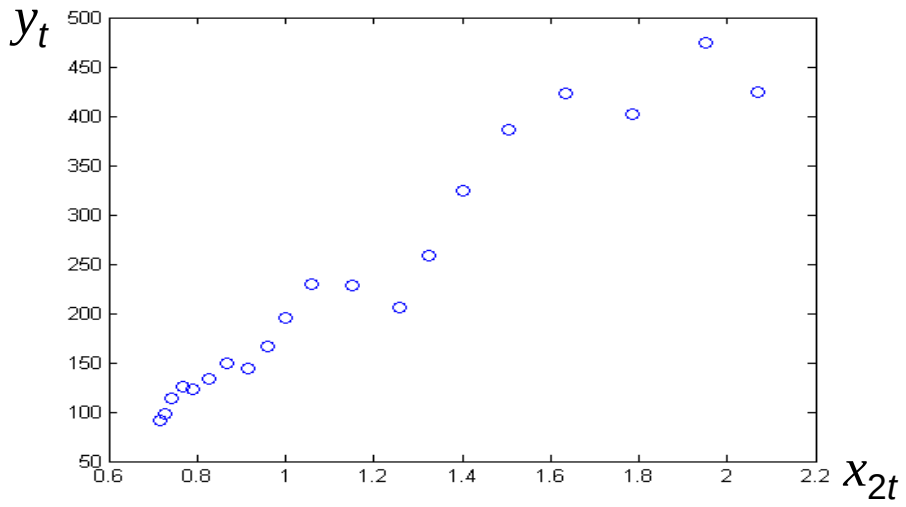
需要诊断并消除数据的自相关性，建立新的模型

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

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

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=0.0000		

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

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

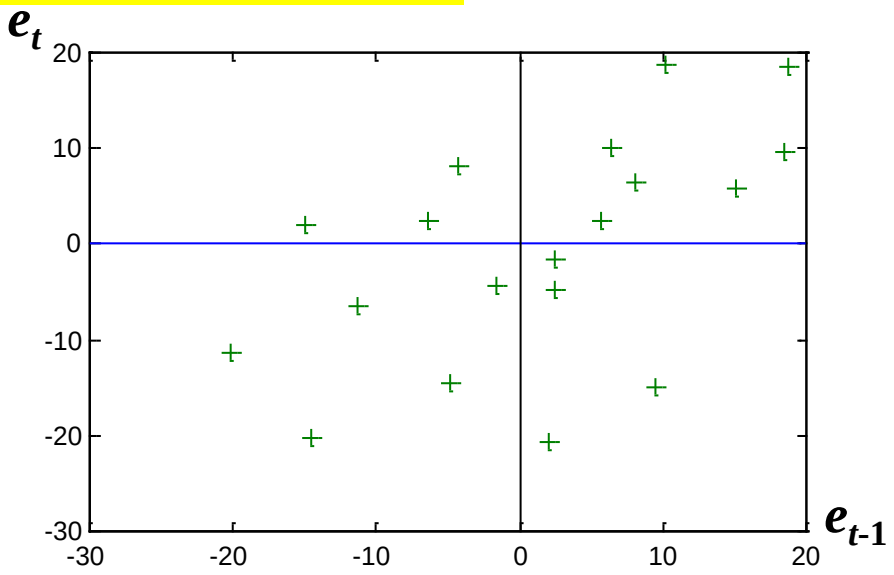
残差诊断法

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

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

*n*较大

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

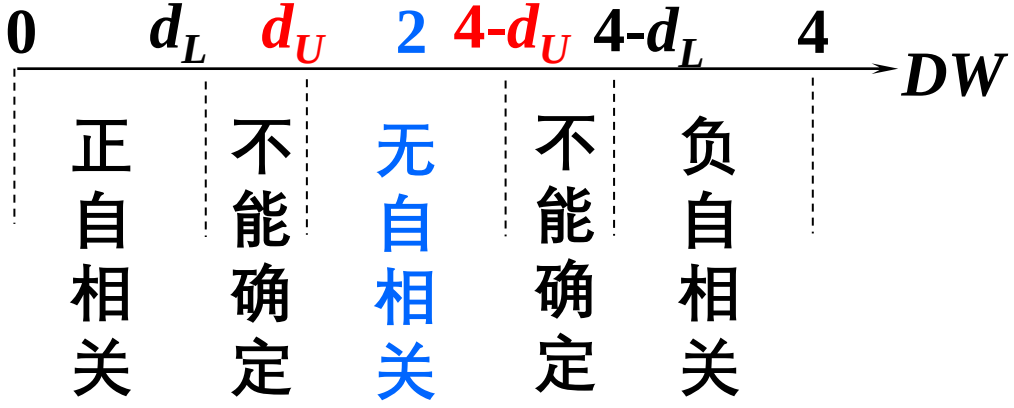
$$=2(1-\hat{\rho})$$

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

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

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

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



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 $DW = 2(1-\hat{\rho}) \Rightarrow \hat{\rho} = 1 - \frac{DW}{2}$

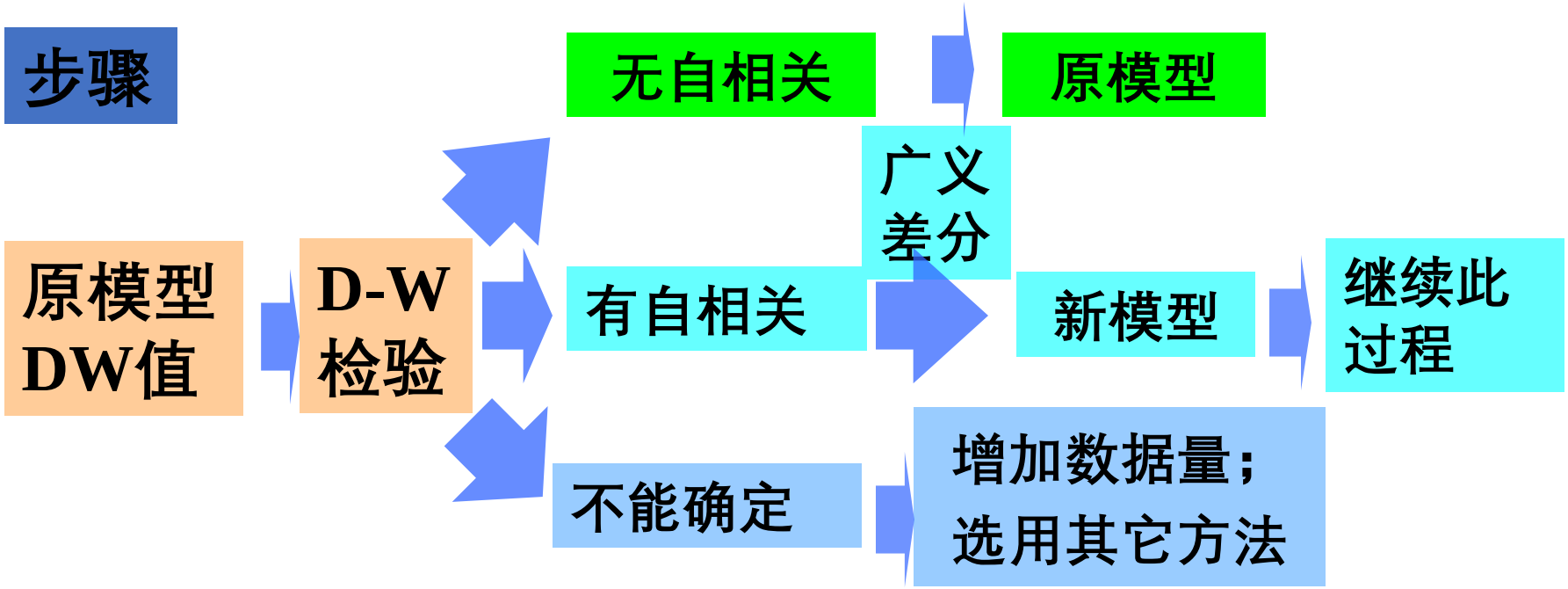
原模型 $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t$

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 $y_t^* = \beta_0^* + \beta_1 x_{1t}^* + \beta_2 x_{2t}^* + u_t \quad \beta_0^* = \beta_0(1-\rho)$

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

原模型残差e_t　DW_{old}=0.8754

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

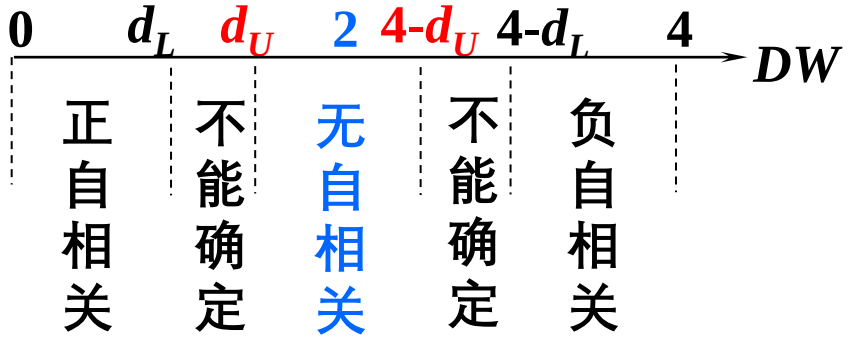
样本容量n=20，回归变量数目k=3，α=0.05

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

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

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

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=0.0000		

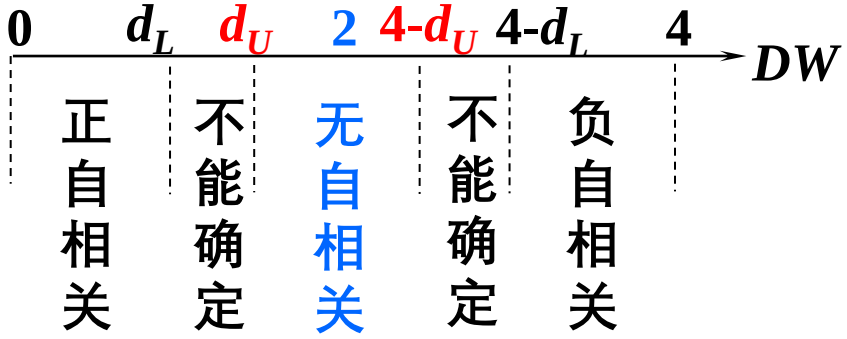
总体效果良好

剩余标准差

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

新模型的自相关性检验

新模型残差e_t ⇒ DW_{new}=1.5751



样本容量n=19，回归变量数目k=3，α=0.05

查表 ⇩

临界值d_L=1.08, d_U=1.53

$d_U < DW_{new} < 4-d_U$

⇩

新模型无自相关性

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

还原为原始变量

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

一阶自回归模型

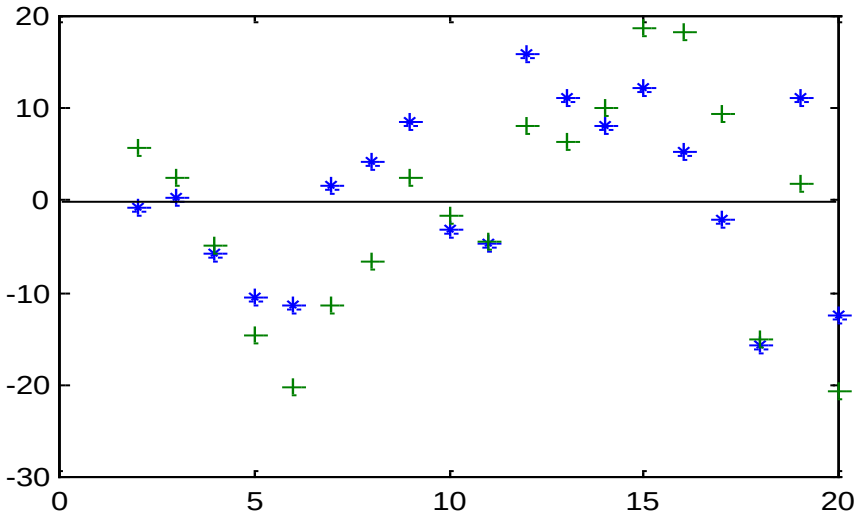
模型结果比较

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

一阶自回归模型

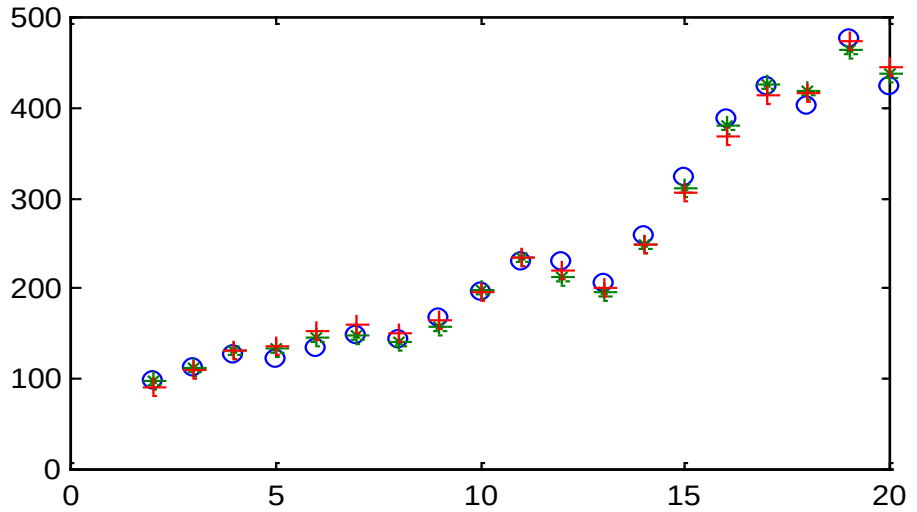
$$\hat{y}_t = 163.4905 + 0.5623y_{t-1} + 0.699x_{1,t} - 0.3930x_{1,t-1} - 1009.0333x_{2,t} + 567.3794x_{2,t-1}$$

残差图比较



新模型 e_t~ *，原模型 e_t~ +

拟合图比较



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

投资额预测

对未来投资额y_t作预测，需先估计出未来的国民生产总值x_{1t}和物价指数 x_{2t}

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	18	401.9	2631.7	1.7842
2	97.4	637.7	0.7277	19	474.9	2954.7	1.9514
3	113.5	691.1	0.7436	20	424.5	3073.0	2.0688

设已知 t=21时， x_{1t}=3312，x_{2t}=2.1938

基本回归模型 $\hat{y}_t = 485.6720$

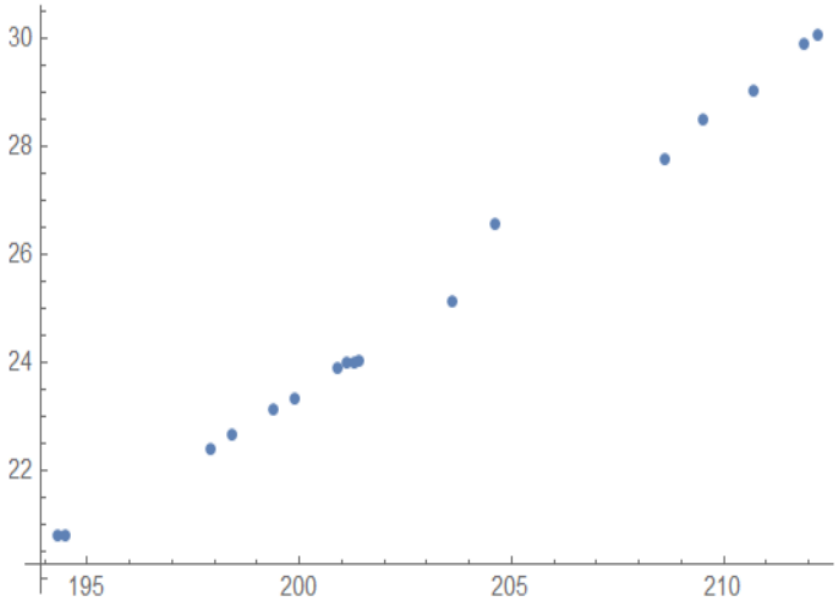
一阶自回归模型 $\hat{y}_t = 469.7638$

$\hat{y}_t$ 较小是由于y_{t-1}=424.5过小所致

练习思考题

例1：（杨启帆《数学建模》§10.5）沸点与气压的实验分析。

#	沸点	气压	#	沸点	气压
1	194.5	20.79	10	201.3	24.01
2	194.3	20.79	11	203.6	25.14
3	197.9	22.40	12	204.6	26.57
4	198.4	22.67	13	209.5	28.49
5	199.4	23.15	14	208.6	27.76
6	199.9	23.35	15	210.7	29.04
7	200.9	23.89	16	211.9	29.88
8	201.1	23.99	17	212.2	30.06
9	201.4	24.02			

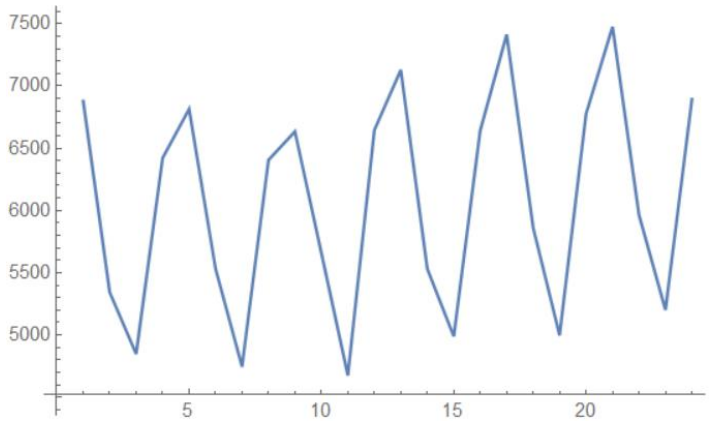


【模型建立】

- 分别作$y_1 = ax + b$，$y_2 = ax^2 + bx + c$，$y_3 = \exp(ax + b)$，$y_4 = \exp(ax^2 + bx + c)$形式的拟合，得$y_2, y_4$的效果差不多。
- 思考：哪个模型好？

例2：（杨启帆《数学建模》§10.5）城市居民用煤量预测。

年份	1季度	2季度	3季度	4季度
1991	6878.4	5343.7	4847.9	6421.9
1992	6815.4	5532.6	4745.6	6406.2
1993	6634.4	5658.5	4674.8	6645.5
1994	7130.2	5532.6	4989.6	6642.3
1995	7413.5	5863.1	4997.4	6776.1
1996	7476.5	5965.5	5202.1	6894.1



【模型建立】

- 居民用煤的主要用途：做饭、取暖。故存在季节性变化。
- 用煤量也与城市人口、经济发展、环境保护等因素有关。
- 模型1：对每个季度的用煤量分别作线性函数拟合。
- 模型2：作拟合$y = ax + b + c_1 \cos\left(\frac{\pi x}{2}\right) + c_2 \sin\left(\frac{\pi x}{2}\right)$。
- 思考：比较模型1和模型2，哪个更好？

例3：（杨启帆《数学建模》§10.6）木材体积的快速折算。

通过测量树干的直径和高度快速估算树干的体积。

#	直径	高度	体积	#	直径	高度	体积	#	直径	高度	体积	#	直径	高度	体积
1	8.3	70	10.3	9	11.1	80	22.6	17	12.9	85	33.8	25	16.3	77	42.6
2	8.6	65	10.3	10	11.2	75	19.9	18	13.3	86	27.4	26	17.3	81	55.4
3	8.8	63	10.2	11	11.3	79	24.2	19	13.7	71	25.7	27	17.5	82	55.7
4	10.5	72	16.4	12	11.4	76	21.0	20	13.8	64	24.9	28	17.9	80	58.3
5	10.7	81	18.8	13	11.4	76	21.4	21	14.0	78	34.5	29	18.0	80	51.5
6	10.8	83	19.7	14	11.7	69	21.3	22	14.2	80	31.7	30	18.0	80	51.0
7	11.0	66	15.6	15	12.0	75	19.1	23	14.5	74	36.3	31	20.6	87	77.0
8	11.0	75	18.2	16	12.9	74	22.2	24	16.0	72	38.3				

【模型建立】

- 模型1：猜测$V = \lambda d^2 h$。算得$\lambda \approx 0.0021$，误差≈ 13.47。
- 模型2：猜测$\ln V = c_0 + c_1 \ln d + c_2 \ln h$。算得$\exp(c_0) \approx 0.0013$，$c_1 \approx 1.98$，$c_2 \approx 1.12$，误差$\approx 13.45$。
- 思考：模型2是否有合理解释？

例4：（杨启帆《数学建模》§10.8）Simpson悖论。

根据下列谋杀案的判决情况分析美国司法制度是否公正?

被告人	被害人	判处死刑	未判死刑
白	白	19	132
白	黑	0	9
黑	白	11	52
黑	黑	6	97

【模型建立】

- 假设是公正的，则每起谋杀案件判处死刑的概率$p \approx \frac{36}{326} \approx 0.11$，应当与被告人和被害人的种族无关。
- 上述4种情形下的比例$p_1 = \frac{19}{151} \approx 0.13$，$p_2 = 0$，$p_3 = \frac{11}{63} \approx 0.17$，$p_4 = \frac{6}{103} \approx 0.06$。显然，$p_2, p_3, p_3$与$p$差别很大。
- 对于每种情形检验“判处死刑的案件数服从概率p的二项分布”。
- 注：“被告人/被害人是白人/黑人”不是随机事件，“是否判决死刑”是随机事件。不应作独立性检验，而应作参数检验。

例5：（CUMCM2012C）脑卒中发病环境因素分析及干预。

脑卒中（俗称脑中风）是目前威胁人类生命的严重疾病之一，它的发生是一个漫长的过程，一旦得病就很难逆转。这种疾病的诱发已经被证实与环境因素，包括气温和湿度之间存在密切的关系。对脑卒中的发病环境因素进行分析，其目的是为了进行疾病的风险评估，对脑卒中高危人群能够及时采取干预措施，也让尚未得病的健康人，或者亚健康人了解自己得脑卒中风险程度，进行自我保护。同时，通过数据模型的建立，掌握疾病发病率的规律，对于卫生行政部门和医疗机构合理调配医务力量、改善就诊治疗环境、配置床位和医疗药物等都具有实际的指导意义。

数据（见Appendix-C1）来源于中国某城市各家医院2007年1月至2010年12月的脑卒中发病病例信息以及相应期间当地的逐日气象资料（Appendix-C2）。请你们根据题目提供的数据，回答以下问题:

1．根据病人基本信息，对发病人群进行统计描述。

2．建立数学模型研究脑卒中发病率与气温、气压、相对湿度间的关系。

3．查阅和搜集文献中有关脑卒中高危人群的重要特征和关键指标，
结合1、2中所得结论，对高危人群提出预警和干预的建议方案。

【模型建立】

- 数据有许多遗漏和错误，格式也不统一。首先需要整理数据。
- 由数据可得，病人中男性占54%，女性占46%，男性比女性易发病。
- 结合各年龄的人口数，发病率（定义为$\frac{\text{病人数}}{\text{人口数}}$）随年龄上升。
- 发病率与湿度的相关系数约-9.5%，空气干燥使发病率上升。
- 发病率与温差的相关系数约3.3%，天气忽冷忽热使发病率上升。

例6：（CUMCM2017C）颜色与物质浓度辨识。

比色法是目前常用的一种检测物质浓度的方法，即把待测物质制备成溶液后滴在特定的白色试纸表面，等其充分反应以后获得一张有颜色的试纸，再把该颜色试纸与一个标准比色卡进行对比，就可以确定待测物质的浓度档位了。由于每个人对颜色的敏感差异和观测误差，使得这一方法在精度上受到很大影响。随着照相技术和颜色分辨率的提高，希望建立颜色读数和物质浓度的数量关系，即只要输入照片中的颜色读数就能够获得待测物质的浓度。试根据附件所提供的有关颜色读数和物质浓度数据完成下列问题：

1. 附件Data1.xls中分别给出了5种物质在不同浓度下的颜色读数，讨论从这5组数据中能否确定颜色读数和物质浓度之间的关系，并给出一些准则来评价这5组数据的优劣。
2. 对附件Data2.xls中的数据，建立颜色读数和物质浓度的数学模型，并给出模型的误差分析。
3. 探讨数据量和颜色维度对模型的影响。

【模型建立】

- 观察数据，推测溶液颜色的R, G, B, H, S值读数是溶液浓度ρ的函数。
- 可选某个主成分指标$X = c_1R + c_2G + c_3B + c_4H + c_5S$，把$\rho$表示成$X$的非线性函数，并把$\rho(X)$的误差作为评价数据优劣的准则。

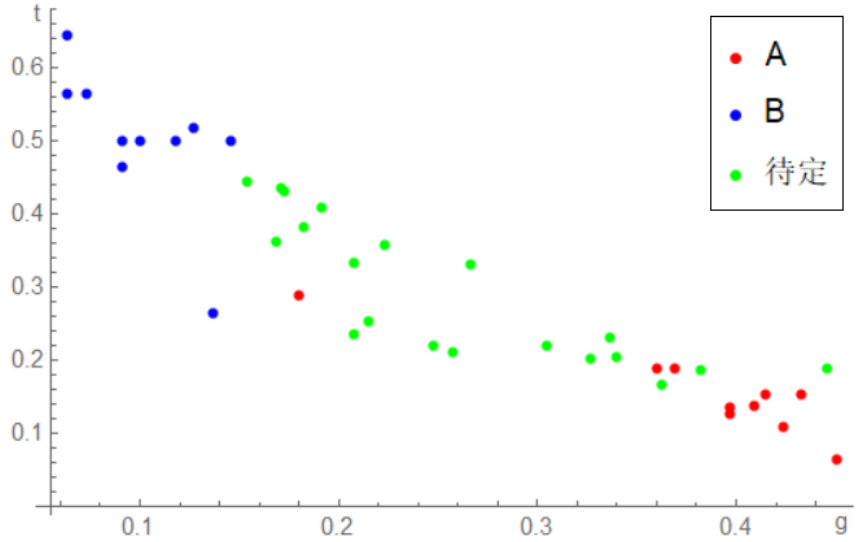
例7：（CUMCM2000A） DNA序列分类。

下面有20个已知类别的人工制造的序列，其中序列标号1~10为A类，11~20为B类。请从中提取特征，构造分类方法，并用这些已知类别的序列，衡量你的方法是否足够好。然后用你认为满意的方法，对另外20个未标明类别的人工序列（标号21~40）进行分类，把结果用序号标明它们的类别。

1. aggcacggaaaaacgggaataacggaggaggacttggcacggcattacacggaggacgaggtaaaggaggcttgtctacggccgga agtgaaaggggatatgaccgcttgg
2. cggaggacaaacgggatggcggtattggaggtggcggactgttcggggaattattcggtttaaacgggacaaggaaggcggctggaac aaccggacggtggcagcaaagga
3. ……
11. gttagatttaacgttttttatggaatttatggaattataaatttaaaaatttatatttttttaggtaagtaatccaacgtttttattacttttttaaaattaaatat ttatt
12. gtttaattactttatcatttaatttaggttttaattttaaatttaatttaggtaagatgaatttggtttttttttaaggtagttatttaattatcgttaaggaaag ttaaa
13. ……

【模型建立】

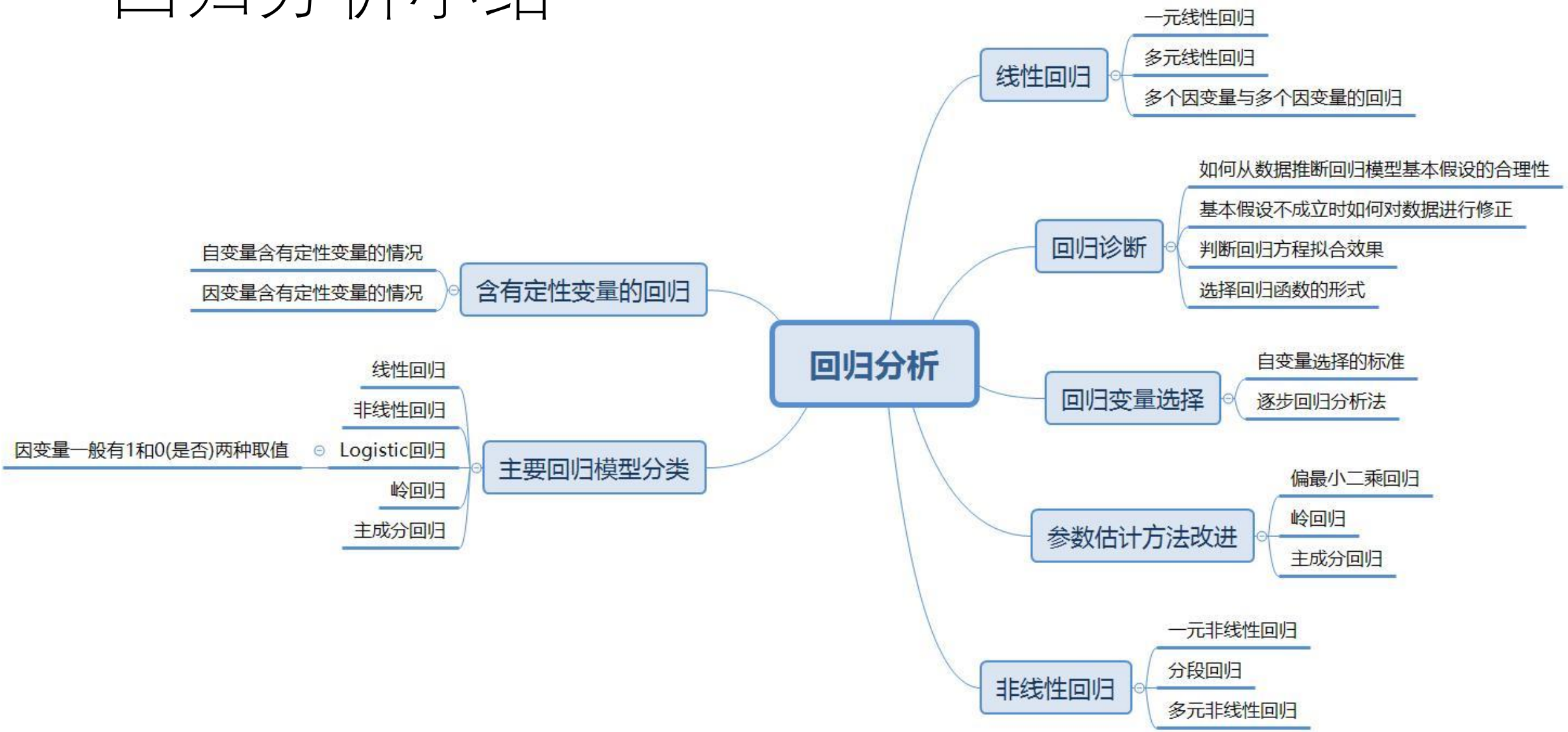
- 经观察后发现，在A类序列中g较多，在B类序列中t较多。
- 故以每个序列中四类碱基的比例(x_1, x_2, x_3, x_4)作为分类指标。
- 当$x_3 \leq 0.16$时，属于A类；否则属于B类。
- 数据量太少，并且缺少生物学知识。
  上述分类方法纯属猜测，无法检验。



统计分析方法

- 统计方法
  - 总体与样本、统计量
- 参数估计方法
- 方差分析法
- 相关分析法
- 显著性检验
- 模型选择

回归分析小结


回归分析
线性回归
一元线性回归
多元线性回归
多个因变量与多个因变量的回归
回归诊断
如何从数据推断回归模型基本假设的合理性
基本假设不成立时如何对数据进行修正
判断回归方程拟合效果
选择回归函数的形式
回归变量选择
自变量选择的标准
逐步回归分析法
参数估计方法改进
偏最小二乘回归
岭回归
主成分回归
非线性回归
一元非线性回归
分段回归
多元非线性回归
含有定性变量的回归
自变量含有定性变量的情况
因变量含有定性变量的情况
主要回归模型分类
线性回归
非线性回归
Logistic回归
因变量一般有1和0(是否)两种取值
岭回归
主成分回归

参考书目

- 何晓群，刘文卿，《应用回归分析》，中国人民大学出版社，2019（第五版）
- 高惠璇，《应用多元统计分析》，北京大学出版社，2014
- 《试验设计与数据分析》

中国科学技术大学

University of Science and Technology of China

谢　谢！