

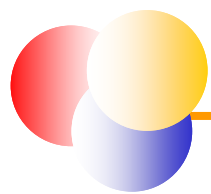
Mathematical modeling

# 第八讲 统计回归模型 (2)

周毓明

zhouyuming@nju.edu.cn

南京大学计算机科学与技术系



# 课程内容

---

1. 数学概念与模型
2. 实际案例与分析
3. 计算机典型应用

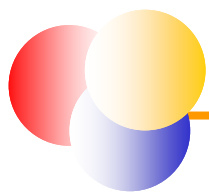


## 2. 实际案例与分析

---

- ① 牙膏的销售量
- ② 程序员的薪金
- ③ 投资额的问题





# 牙膏的销售量

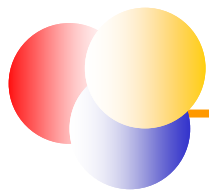
## 问题

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



# 牙膏的销售量

## 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

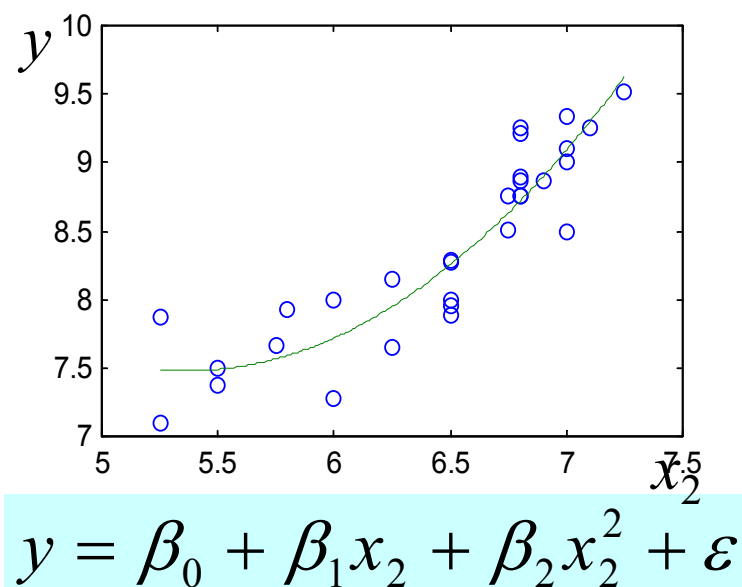
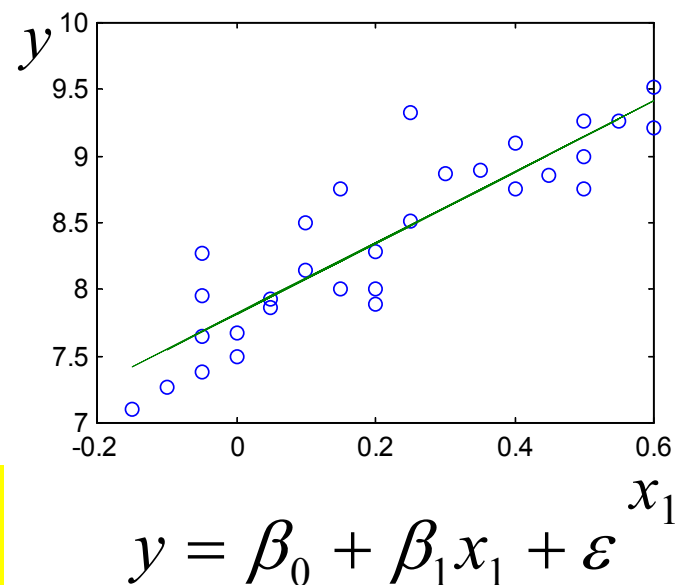
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$

$y$  ~ 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

$\varepsilon$  ~ 随机误差 (均值为零的正态分布随机变量)



## 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  $\beta$

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

输入

$y \sim n$  维数据向量

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

`alpha`(置信水平,0.05)

输出

$b \sim \beta$  的估计值

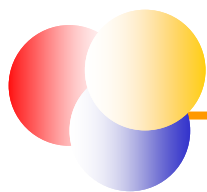
`bint`  $\sim b$  的置信区间

$r \sim$  残差向量  $y - xb$

`rint`  $\sim r$  的置信区间

参数	参数估计值	置信区间
$\beta_0$	17.3244	[5.7282 28.9206]
$\beta_1$	1.3070	[0.6829 1.9311]
$\beta_2$	-3.6956	[-7.4989 0.1077]
$\beta_3$	0.3486	[0.0379 0.6594]
$R^2=0.9054 \quad F=82.9409 \quad p=0.0000$		

`Stats`  $\sim$   
检验统计量  
 $R^2, F, p$



# 牙膏的销售量

## 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
$\beta_0$	17.3244	[5.7282 28.9206]
$\beta_1$	1.3070	[0.6829 1.9311 ]
$\beta_2$	-3.6956	[-7.4989 0.1077 ]
$\beta_3$	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 $\alpha=0.05$

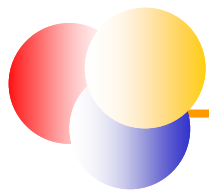
模型从整体上看成立

$\beta_2$ 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



# 牙膏的销售量



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  $7.8320 \times 3.7 \approx 29$  (百万元) 以上



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

模型改进

$x_1$ 和 $x_2$ 对 $y$ 的  
影响独立

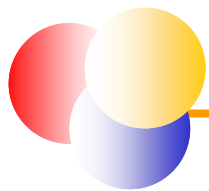


$x_1$ 和 $x_2$ 对 $y$   
的影响有  
交互作用

参数	参数估计值	置信区间
$\beta_0$	17.3244	[5.7282 28.9206]
$\beta_1$	1.3070	[0.6829 1.9311]
$\beta_2$	-3.6956	[-7.4989 0.1077]
$\beta_3$	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
$\beta_0$	29.1133	[13.7013 44.5252]
$\beta_1$	11.1342	[1.9778 20.2906]
$\beta_2$	-7.6080	[-12.6932 -2.5228]
$\beta_3$	0.6712	[0.2538 1.0887]
$\beta_4$	-1.4777	[-2.8518 -0.1037]
$R^2=0.9209$ $F=72.7771$ $p=0.0000$		



# 牙膏的销售量



## 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

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

$$\hat{y} = 8.2933 \text{ (百万支)}$$

区间 [7.8230, 8.7636]

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

$$\hat{y} = 8.3272 \text{ (百万支)}$$

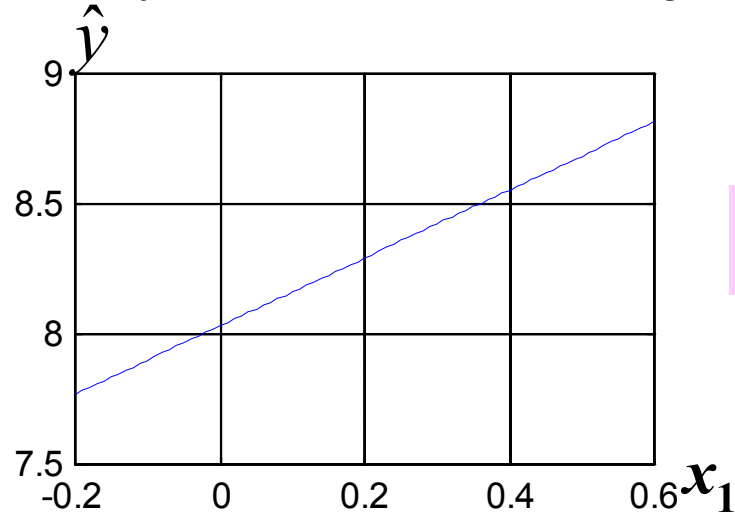
区间 [7.8953, 8.7592]

$\hat{y}$  略有增加

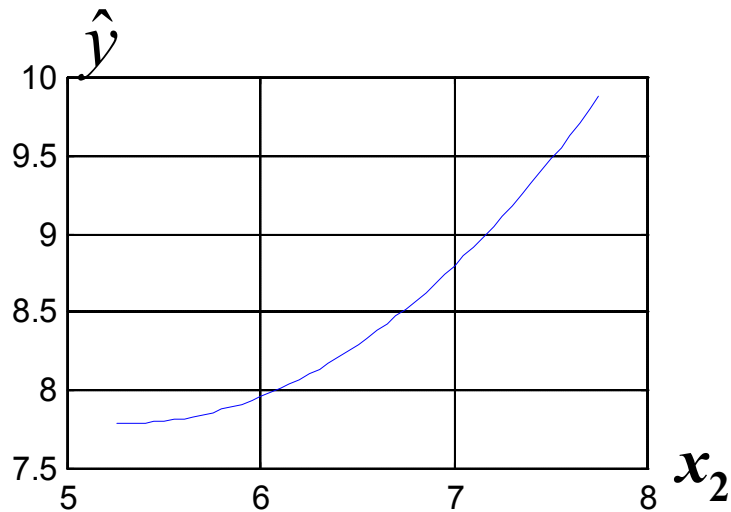
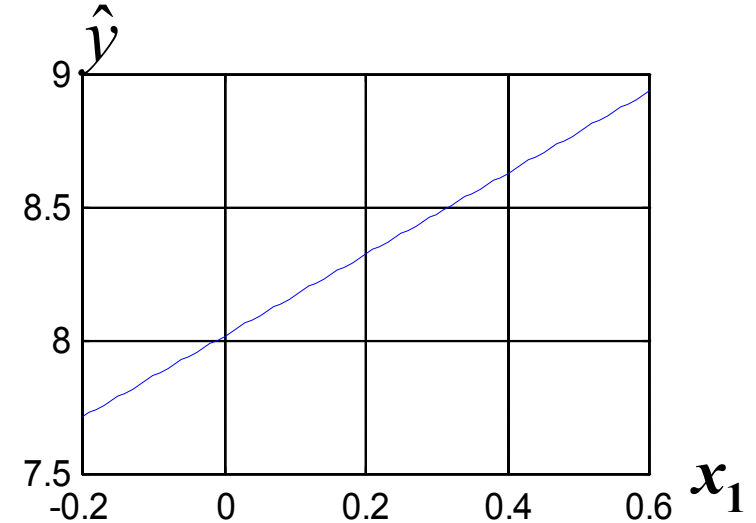
预测区间长度更短

## 两模型 $\hat{y}$ 与 $x_1, x_2$ 关系的比较

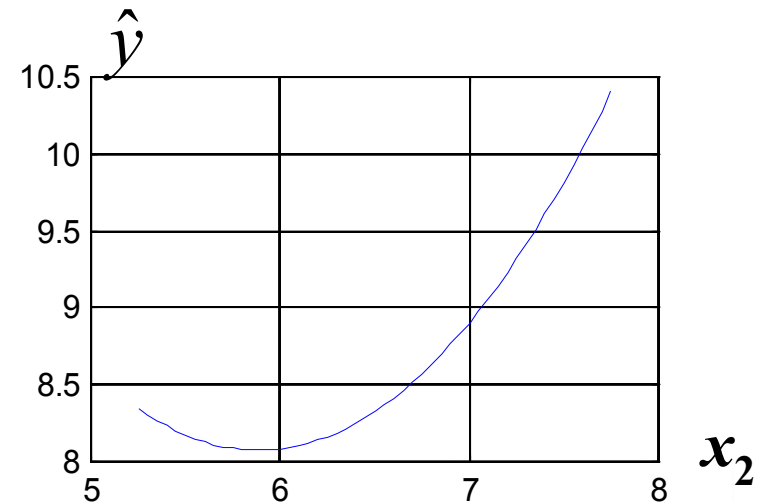
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 \quad \hat{y} = \beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 x_2$$



$x_2=6.5$



$x_1=0.2$



## 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  $x_1=0.1$

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

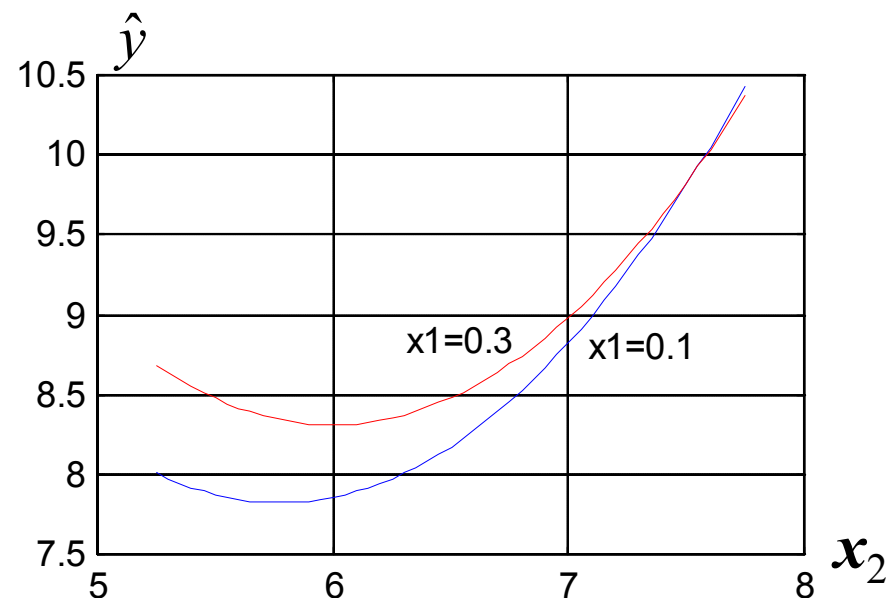
价格差  $x_1=0.3$

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

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

价格优势会使销售量增加

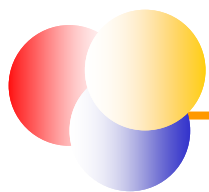
加大广告投入使销售量增加  
( $x_2$  大于6百万元)



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

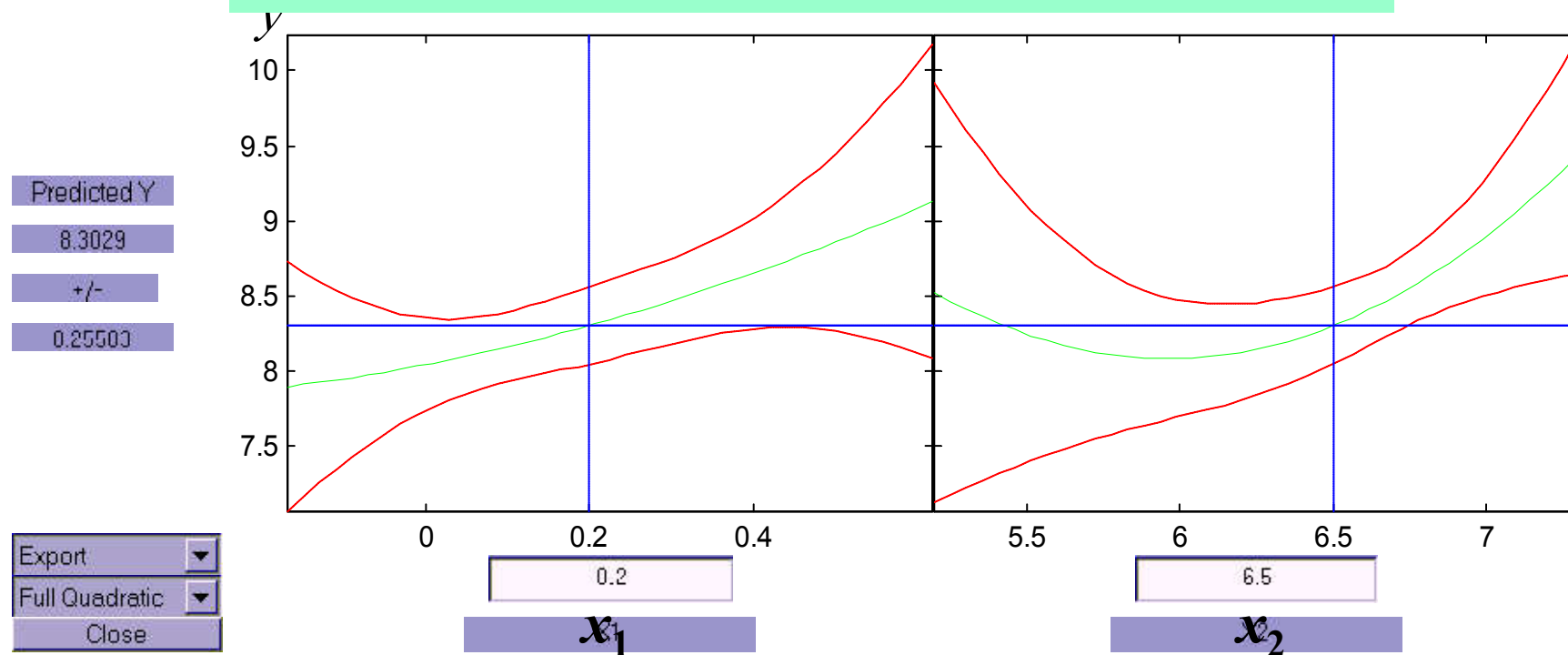


# 牙膏的销售量

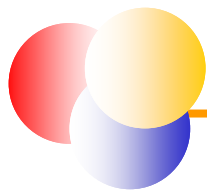
## 完全二次多项式模型

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

MATLAB中有命令rstool直接求解



从输出 **Export** 可得  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$  <sup>13</sup>



# 程序员的薪金

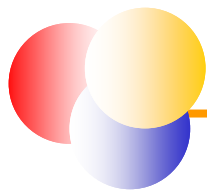
建立模型研究薪金与资历、管理责任、教育程度的关系

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



# 程序员的薪金

分析与假设

$y \sim$  薪金,  $x_1 \sim$  资历 (年)

$x_2 = 1 \sim$  管理人员,  $x_2 = 0 \sim$  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

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

$a_0, a_1, \dots, a_4$  是待估计的回归系数,  $\varepsilon$  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
$a_0$	11032	[ 10258 11807 ]
$a_1$	546	[ 484 608 ]
$a_2$	6883	[ 6248 7517 ]
$a_3$	-2994	[ -3826 -2162 ]
$a_4$	148	[ -636 931 ]
$R^2=0.957 \quad F=226 \quad p=0.000$		

资历增加1年薪  
金增长546

管理人员薪金多  
6883

中学程度薪金比研  
究生的少2994

大学程度薪金比研  
究生多148

$R^2, F, p \rightarrow$  模型整体上可用

$x_1 \sim$  资历(年)

$x_2 = 1 \sim$  管理,  $x_2 = 0 \sim$  非管理

中学:  $x_3 = 1, x_4 = 0$ ;

大学:  $x_3 = 0, x_4 = 1$ ;

研究生:  $x_3 = 0, x_4 = 0$ .

$a_4$  置信区间包含零点,  
解释不可靠!



## 结果分析

## 残差分析方法

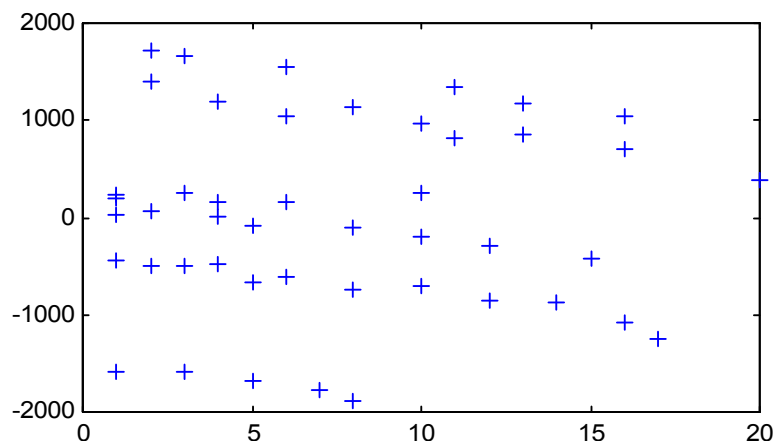
$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \hat{a}_3 x_3 + \hat{a}_4 x_4$$

残差  $e = y - \hat{y}$

## 管理与教育的组合

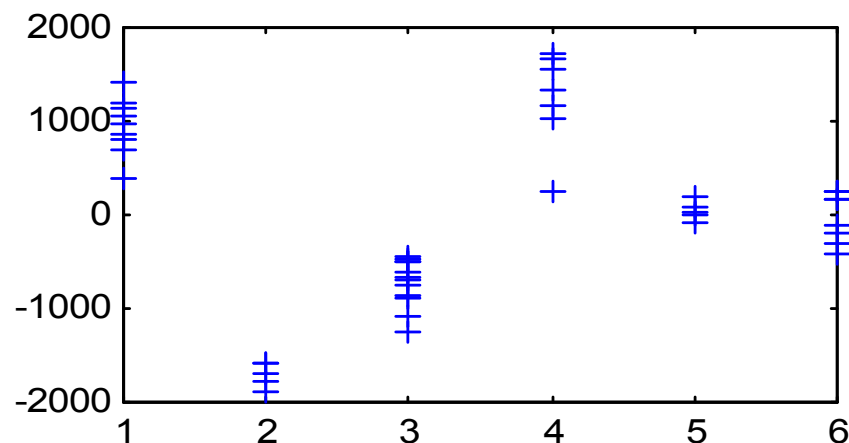
组合	1	2	3	4	5	6
管理	0	1	0	1	0	1
教育	1	1	2	2	3	3

### $e$ 与资历 $x_1$ 的关系



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。

### $e$ 与管理—教育组合的关系



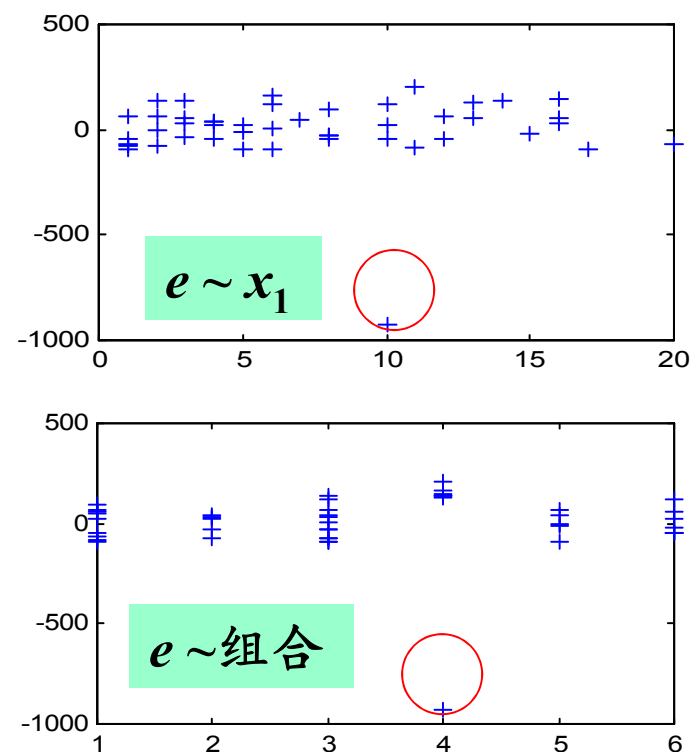
残差全为正，或全为负，管理—教育组合处理不当

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

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

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

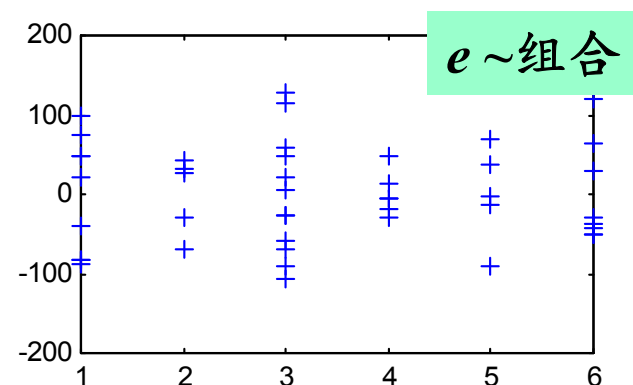
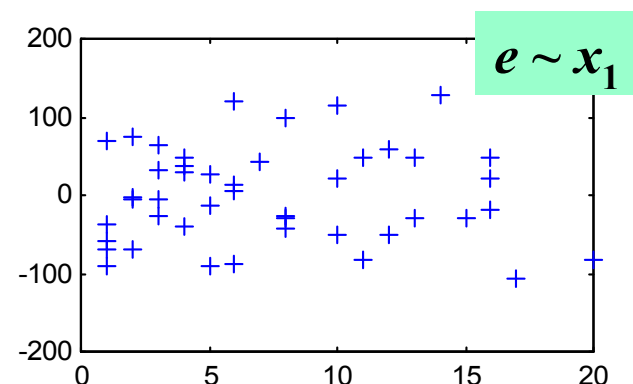
## 去掉异常数据后的结果

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  $\rightarrow$  0.999  $\rightarrow$  0.9998

$F$ : 226  $\rightarrow$  554  $\rightarrow$  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)

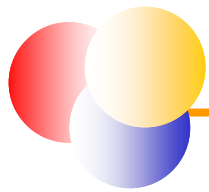
$x_1=0$ ;  $x_2=1$ ~ 管理,  $x_2=0$ ~ 非管理

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



# 程序员的薪金

## 软件开发人员的薪金

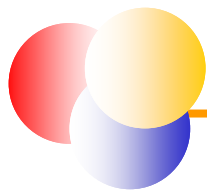
对定性因素(如管理、教育), 可以引入0-1变量处理,  
0-1变量的个数应比定性因素的水平少1

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





# 投资额的问题

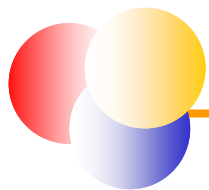
问题

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



# 投资额的问题

## 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

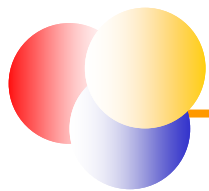
以时间为序的数据，称为时间序列

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
...	...	...	...	...	...	...	23 ...

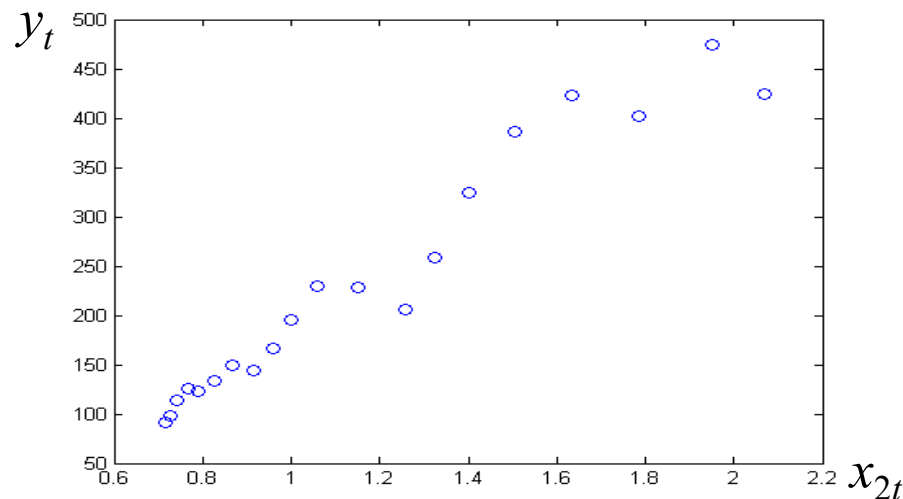
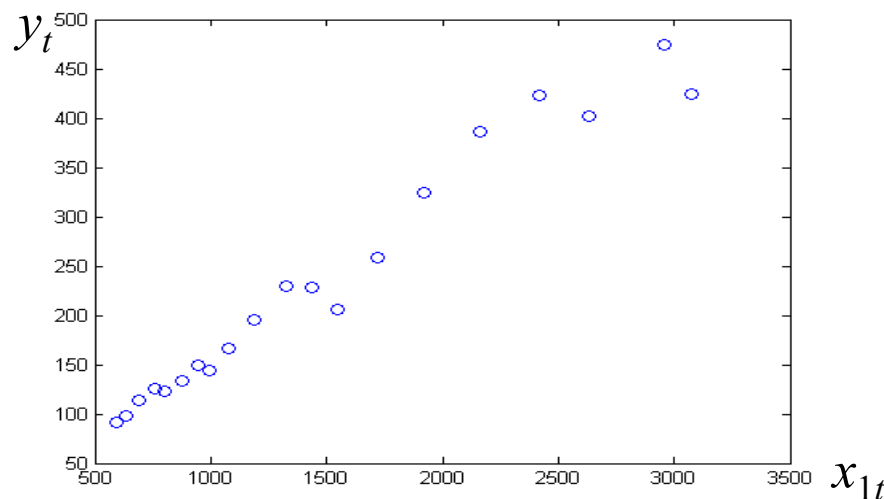


# 投资额的问题



## 基本回归模型

$t \sim$  年份,  $y_t \sim$  投资额,  $x_{1t} \sim$  GNP,  $x_{2t} \sim$  物价指数



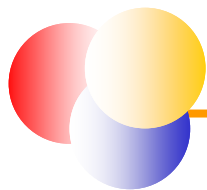
投资额与 GNP 及物价指数间均有很强的线性关系

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t \quad \beta_0, \beta_1, \beta_2 \sim \text{回归系数}$$

$\varepsilon_t \sim$  对  $t$  相互独立的零均值正态随机变量







# 投资额的问题

## 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
$\beta_0$	322.7250	[224.3386 421.1114]
$\beta_1$	0.6185	[0.4773 0.7596]
$\beta_2$	-859.4790	[-1121.4757 -597.4823 ]
$R^2= 0.9908 \quad F= 919.8529 \quad p=0.0000$		

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

剩余标准差  
 $s=12.7164$

模型优点

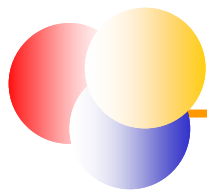
$R^2=0.9908$ ，拟合度高

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





# 投资额的问题

## 自相关性的定性诊断

模型残差  $e_t = y_t - \hat{y}_t$

$e_t$  为随机误差  $\varepsilon_t$  的估计值

在MATLAB工作区中输出

作残差  $e_t \sim e_{t-1}$  散点图

大部分点落在第1,3象限

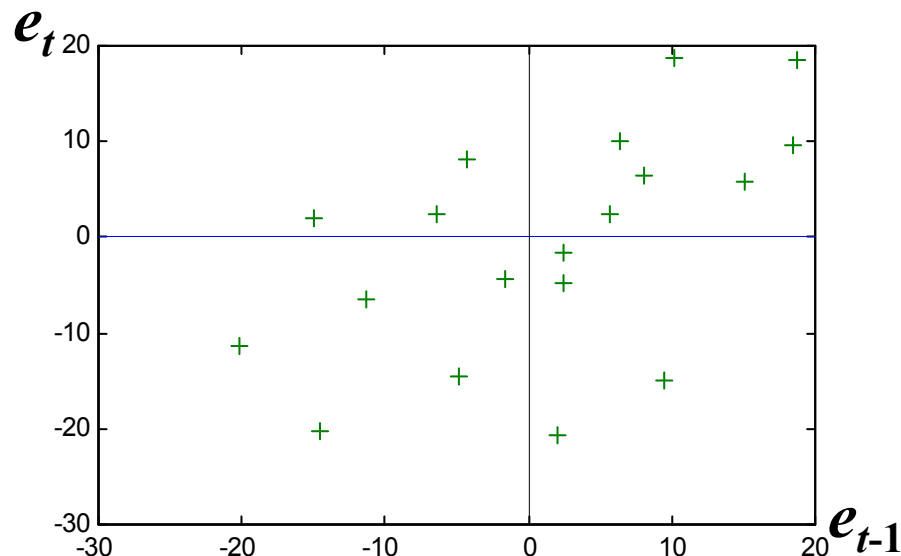
大部分点落在第2,4象限

自相关性直观判断

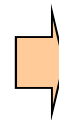


基本回归模型的随机误差项  
 $\varepsilon_t$  存在正的自相关

## 残差诊断法

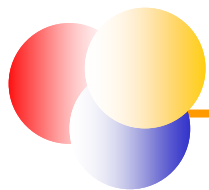


$\varepsilon_t$  存在正的自相关



$\varepsilon_t$  存在负的自相关





# 投资额的问题

自回归性的定量诊断

D-W检验

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

$\beta_0, \beta_1, \beta_2 \sim$  回归系数

$\rho \sim$  自相关系数

$|\rho| \leq 1$

$u_t \sim$  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  $\rho$

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

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

$$\approx 2 \left[ 1 - \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=2}^n e_t^2} \right] \quad n \text{较大}$$

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

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

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

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

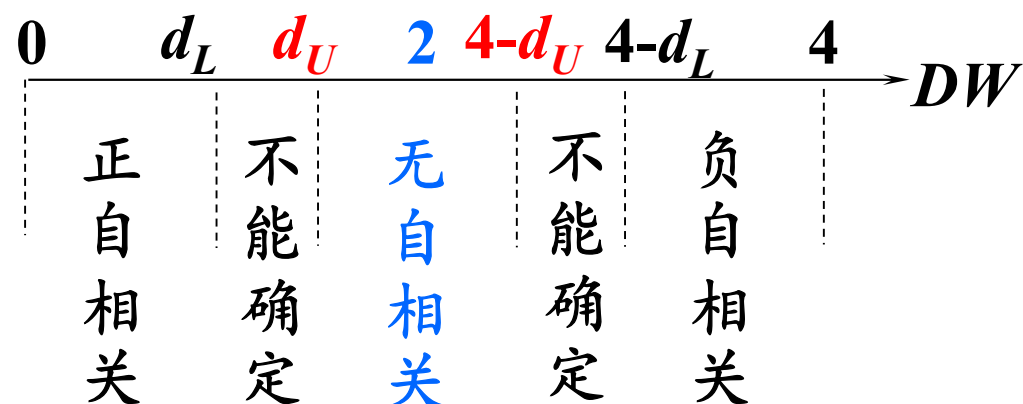
$$\hat{\rho} = -1 \rightarrow DW = 4$$

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

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

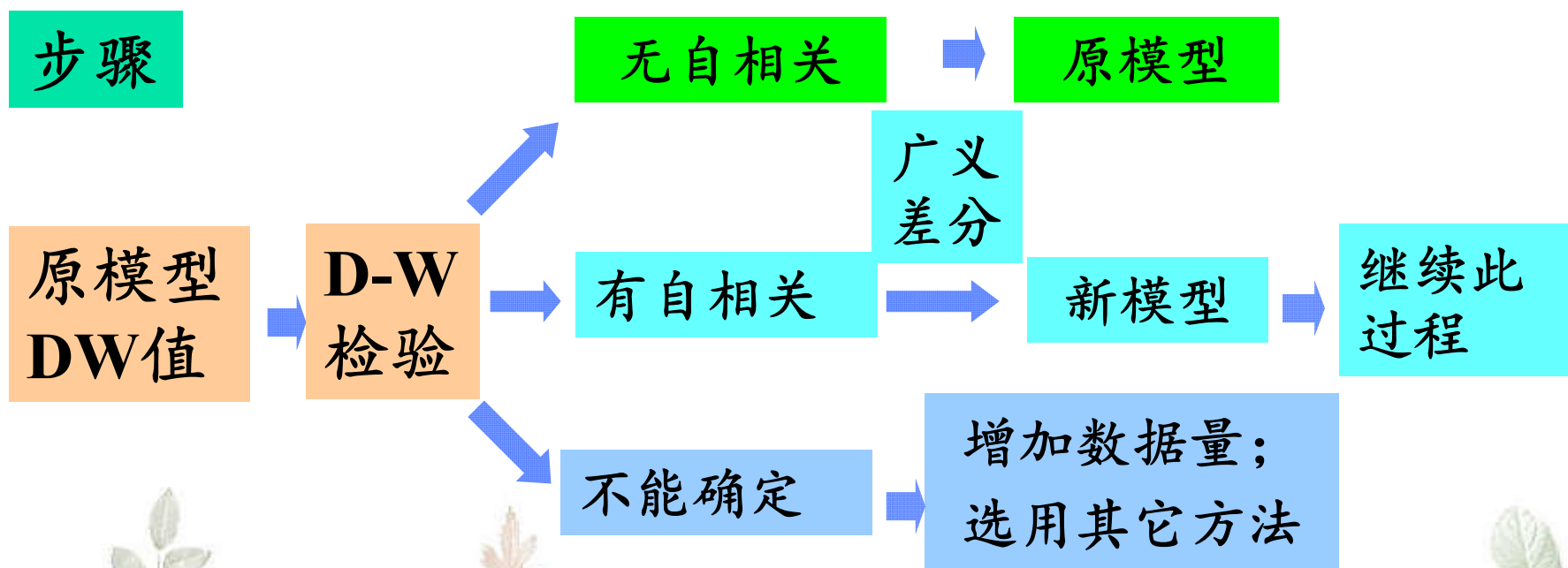
广义差分变换  $DW = 2(1 - \hat{\rho}) \iff \hat{\rho} = 1 - \frac{DW}{2}$

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

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



# 投资额新模型的建立

原模型  
残差  $e_t$   $DW_{old} = 0.8754$

样本容量  $n=20$ , 回归  
变量数目  $k=3$ ,  $\alpha=0.05$

查表 

临界值  $d_L=1.10$ ,  $d_U=1.54$

作变换

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

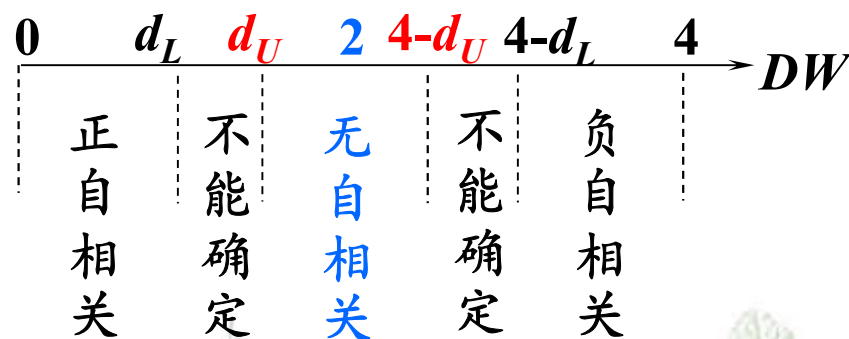
$$x_{it}^* = x_{it} - 0.5623x_{i,t-1}, \quad i=1,2$$

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

$$DW_{old} < d_L$$

原模型有  
正自相关

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



## 投资额新模型的建立

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

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

由数据  $y_t^*, x_{1t}^*, x_{2t}^*$  估计系数  $\beta_0^*, \beta_1, \beta_2$

参数	参数估计值	置信区间
$\beta_0^*$	163.4905	[1265.4592 2005.2178]
$\beta_1$	0.6990	[0.5751 0.8247]
$\beta_2$	-1009.0333	[-1235.9392 -782.1274]
$R^2=0.9772 \quad F=342.8988 \quad p=0.0000$		

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  $e_t$



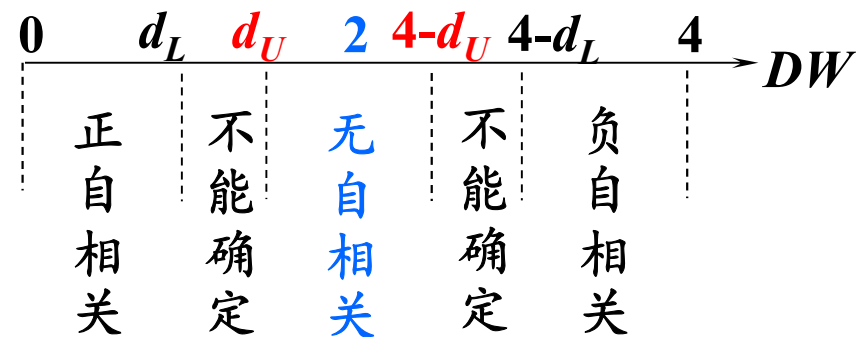
$$DW_{new} = 1.5751$$

样本容量  $n=19$ ，回归  
变量数目  $k=3$ ， $\alpha=0.05$

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

还原为  
原始变量

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

一阶自回归模型

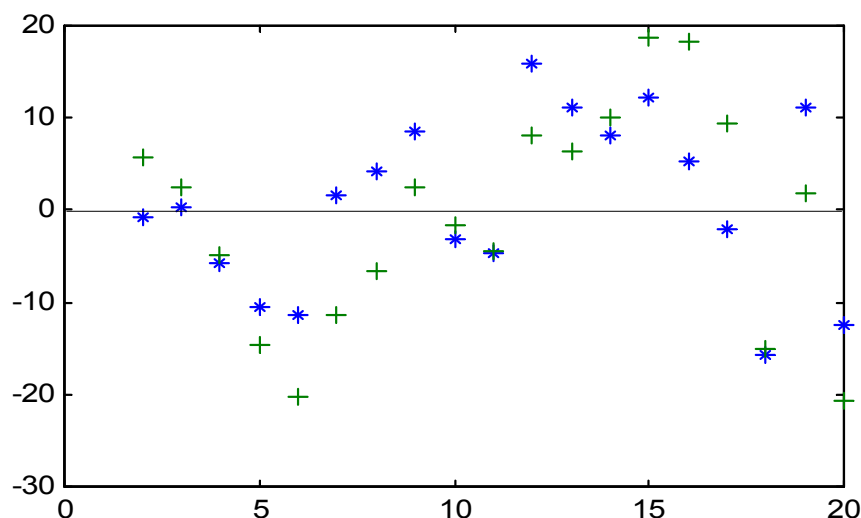


## 模型结果比较

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

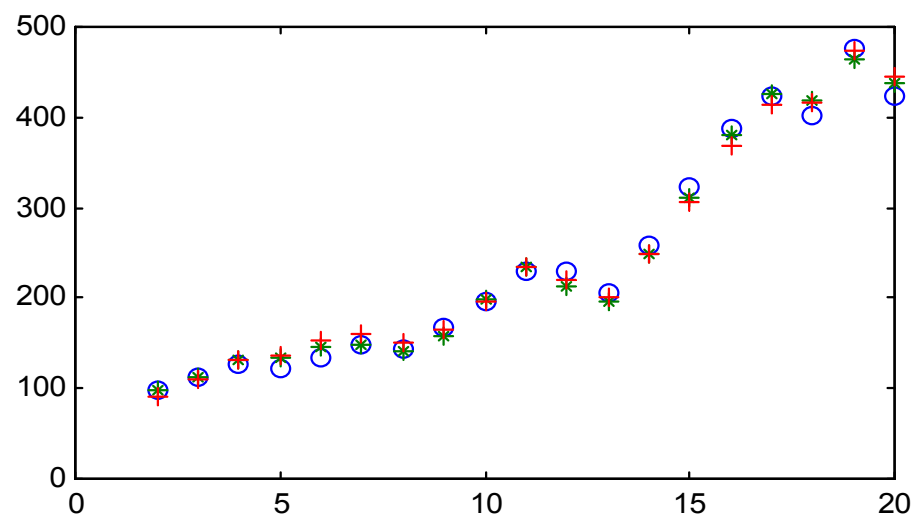
一阶自回归模型  $\hat{y}_t = 163.4905 + 0.5623y_{t-1} + 0.699x_{1,t} - 0.3930x_{1,t-1} - 1009.0333x_{2,t} + 567.3794x_{2,t-1}$

残差图比较



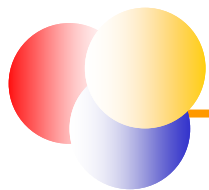
新模型  $e_t \sim *$ , 原模型  $e_t \sim +$

拟合图比较



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

一阶自回归模型残差  $e_t$  比基本回归模型要小



# 投资额的问题

## 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

Thanks for your time and  
attention!



# 一元线性回归分析的应用：预测问题

一、 $\hat{Y}_0$ 是条件均值 $E(Y | X=X_0)$ 或个值 $Y_0$ 的一个无偏估计

二、总体条件均值与个值预测值的置信区间



对于一元线性回归模型

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

给定样本以外的解释变量的观测值 $X_0$ ，可以得到被解释变量的预测值 $\hat{Y}_0$ ，可以此作为其条件均值 $E(Y|X=X_0)$ 或个别值 $Y_0$ 的一个近似估计。

注意：

严格地说，这只是被解释变量的预测值的估计值，而不是预测值。

原因：（1）参数估计量不确定；

（2）随机项的影响



## 一、 $\hat{Y}_0$ 是条件均值 $E(Y|X=X_0)$ 或个值 $Y_0$ 的一个无偏估计

对总体回归函数 $E(Y|X)=\beta_0+\beta_1X$ ,  $X=X_0$ 时

$$E(Y|X=X_0)=\beta_0+\beta_1X_0$$

通过样本回归函数 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$ , 求得的拟合值为

$$\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1X_0$$

$$\text{于是 } E(\hat{Y}_0)=E(\hat{\beta}_0+\hat{\beta}_1X_0)=E(\hat{\beta}_0)+X_0E(\hat{\beta}_1)=\beta_0+\beta_1X_0$$

可见,  $\hat{Y}_0$ 是条件均值 $E(Y|X=X_0)$ 的无偏估计。



对总体回归模型 $Y = \beta_0 + \beta_1 X + \mu$ ，当 $X = X_0$ 时

$$Y_0 = \beta_0 + \beta_1 X_0 + \mu$$

于是

$$E(Y_0) = E(\beta_0 + \beta_1 X_0 + \mu) = \beta_0 + \beta_1 X_0 + E(\mu) = \beta_0 + \beta_1 X_0$$

而通过样本回归函数 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ ，求得拟合值

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

的期望为

$$E(\hat{Y}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 X_0) = E(\hat{\beta}_0) + X_0 E(\hat{\beta}_1) = \beta_0 + \beta_1 X_0$$

$\hat{Y}_0$ 是个值 $Y_0$ 的无偏估计。



## 二、总体条件均值与个值预测值的置信区间

### 1、总体均值预测值的置信区间

由于  $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2}) \quad \hat{\beta}_0 \sim N(\beta_0, \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2)$$

于是  $E(\hat{Y}_0) = E(\hat{\beta}_0) + X_0 E(\hat{\beta}_1) = \beta_0 + \beta_1 X_0$

$$Var(\hat{Y}_0) = Var(\hat{\beta}_0) + 2X_0 Cov(\hat{\beta}_0, \hat{\beta}_1) + X_0^2 Var(\hat{\beta}_1)$$

可以证明  $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \bar{X} / \sum x_i^2$





因此

$$\begin{aligned} \text{Var}(\hat{Y}_0) &= \frac{\sigma^2 \sum X_i^2}{n \sum x_i^2} - \frac{2X_0 \bar{X} \sigma^2}{\sum x_i^2} + \frac{X_0^2 \sigma^2}{\sum x_i^2} \\ &= \frac{\sigma^2}{\sum x_i^2} \left( \frac{\sum X_i^2 - n\bar{X}^2}{n} + \bar{X}^2 - 2X_0 \bar{X} + X_0^2 \right) \\ &= \frac{\sigma^2}{\sum x_i^2} \left( \frac{\sum x_i^2}{n} + (X_0 - \bar{X})^2 \right) = \sigma^2 \left( \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right) \end{aligned}$$

故

$$\hat{Y}_0 \sim N(\beta_0 + \beta_1 X_0, \sigma^2 (\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}))$$

将未知的 $\sigma^2$ 代以它的无偏估计量  $\hat{\sigma}^2$ ，可构造 **t统计量**

$$t = \frac{\hat{Y}_0 - (\beta_0 + \beta_1 X_0)}{S_{\hat{Y}_0}} \sim t(n-2) \quad \text{其中} \quad S_{\hat{Y}_0} = \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right)}$$

于是，在 $1-\alpha$ 的置信度下，**总体均值 $E(Y|X_0)$ 的置信区间为**

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0} < E(Y | X_0) < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0}$$

## 2、总体个值预测值的预测区间

由  $Y_0 = \beta_0 + \beta_1 X_0 + \mu$  知：

$$Y_0 \sim N(\beta_0 + \beta_1 X_0, \sigma^2)$$

于是  $\hat{Y}_0 - Y_0 \sim N(0, \sigma^2(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}))$

将未知的 $\sigma^2$ 代以它的无偏估计量 $\hat{\sigma}^2$ ，可构造t统计量

$$t = \frac{\hat{Y}_0 - Y_0}{S_{\hat{Y}_0 - Y_0}} \sim t(n-2)$$

式中：
$$S_{\hat{Y}_0 - Y_0} = \sqrt{\hat{\sigma}^2(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2})}$$

从而在 $1-\alpha$ 的置信度下， $Y_0$ 的置信区间为

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0 - Y_0} < Y_0 < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0 - Y_0}$$



在上述**收入-消费支出**例中，得到的样本回归函数为

$$\hat{Y}_i = -103.172 + 0.777 X_i$$

则在  $X_0=1000$ 处，  $\hat{Y}_0 = -103.172 + 0.777 \times 1000 = 673.84$

而

$$Var(\hat{Y}_0) = 13402 \left[ \frac{1}{10} + \frac{(1000 - 2150)^2}{7425000} \right] = 3727.29$$

$$S(\hat{Y}_0) = 61.05$$

因此，**总体均值** $E(Y|X=1000)$ 的95%的置信区间为：

$$673.84 - 2.306 \times 61.05 < E(Y|X=1000) < 673.84 + 2.306 \times 61.05$$

或  $(533.05, 814.62)$



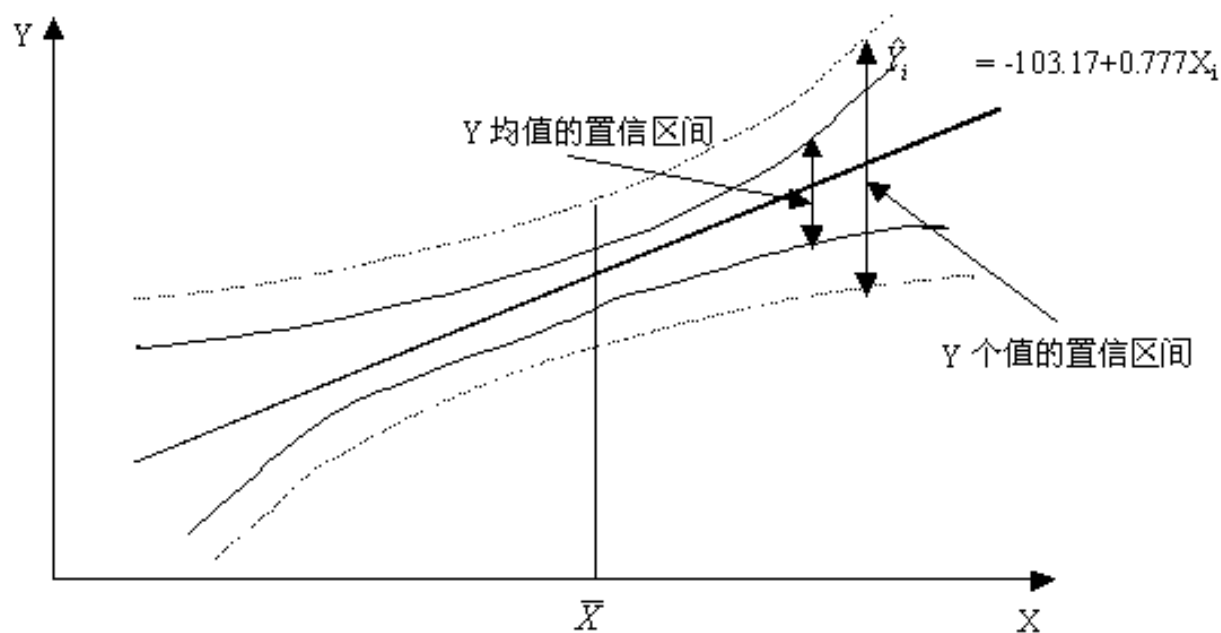
同样地，对于Y在X=1000的**个体值**，其95%的置信区间为：

$$673.84 - 2.306 \times 130.88 < Y_{x=1000} < 673.84 + 2.306 \times 130.88$$

或

$$(372.03, 975.65)$$

- 总体回归函数的**置信带（域）**（confidence band）
- 个体的**置信带（域）**



对于Y的总体均值 $E(Y|X)$ 与个体值的预测区间（置信区间）：

（1）样本容量 $n$ 越大，预测精度越高，反之预测精度越低；

（2）样本容量一定时，置信带的宽度当在 $X$ 均值处最小，其附近进行预测（插值预测）精度越大； $X$ 越远离其均值，置信带越宽，预测可信度下降。

