

- 1 牙膏的销售量
- 2 软件开发人员的薪金
- 3 酶促反应
投资额模型

第十章统计回归模型

① 1 牙膏的销售量

② 2 软件开发人员的薪金

③ 3 酶促反应

④ 投资额模型

数学建模的基本方法

机理分析和测试分析

- 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

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

问题

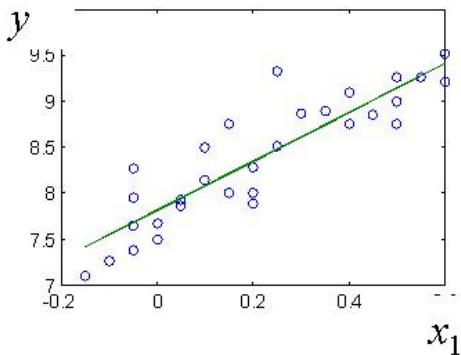
销售 周期	本公司价 格(元)	其它厂家 价格(元)	广告费用 (百万元)	价格差 (元)	销售量 (百万支)
1	3.85	3.80	5.50	-0.05	7.38
2	3.75	4.00	6.75	0.25	8.51
...
29	3.80	3.85	5.80	0.05	7.93
30	3.70	4.25	6.80	0.55	9.26

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

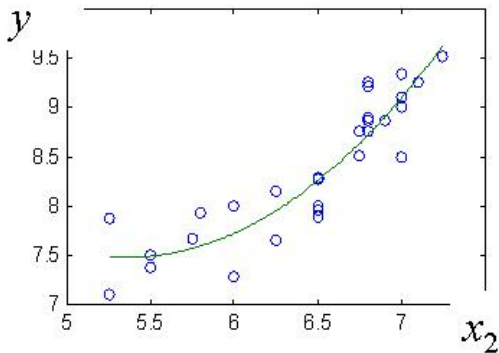
- 1 牙膏的销售量
- 2 软件开发人员的薪金
- 3 酶促反应
- 投资额模型

Model



- 1 牙膏的销售量
 - 2 软件开发人员的薪金
 - 3 酶促反应
- 投资额模型

Model



Matlab 求解

- data=load('yagaoData.txt');
y=data(:,5); % 销售量
x1=data(:,4); % 价格差
x2=data(:,3); % 广告费
- x=[ones(30,1),x1,x2,x2.^2];
- [b,bint,r,rint,stats]=regress(y,x,0.05);

Model

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]
$R^2=0.9054$ $F=82.9409$ $p=0.0000$		

结果分析

- 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
- 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

- 价格差 x_1 =其它厂家价格 x_3 -本公司价格 x_4
- 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
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 = 8.2933$
- 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 $7.8320 \times 3.7 \simeq 29$ (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
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2 + \varepsilon$

- 1 牙膏的销售量
- 2 软件开发人员的薪金
- 3 酶促反应
- 投资额模型

模型二计算结果

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,
- 模型一: $\hat{y} = 8.2933$ (百万支), 区间 $[7.8230, 8.7636]$
- 模型二: $\hat{y} = 8.3272$ (百万支), 区间 $[7.8953, 8.7592]$
- 预测略有增加, 预测区间长度更短

交互作用影响的讨论

- 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 $x_1=0.3$

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

●

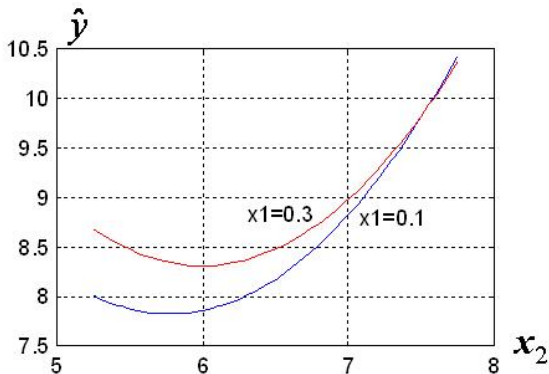
$$x_2 < 7.5357$$

$$\hat{y}|_{x_1=0.3} > \hat{y}|_{x_1=0.1}$$

- 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)

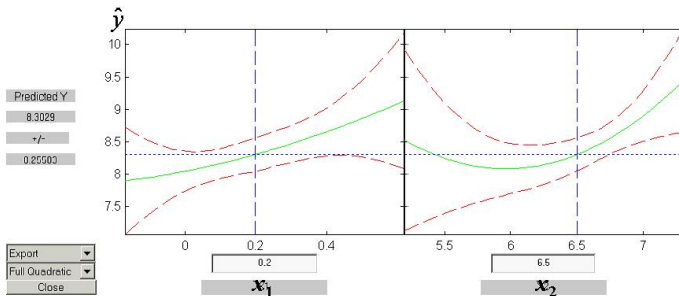
- 1 牙膏的销售量
- 2 软件开发人员的薪金
- 3 酶促反应
- 投资额模型

Model



- 1 牙膏的销售量
 - 2 软件开发人员的薪金
 - 3 酶促反应
- 投资额模型

Model



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见后面的图
- 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

Model

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

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

-

$$x_3 = \begin{cases} 1, & \text{MiddleSchool} \\ 0, & \text{otherwise} \end{cases}$$

-

$$x_4 = \begin{cases} 1, & \text{University} \\ 0, & \text{otherwise} \end{cases}$$

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

分析与假设

- 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, \dots, a_4 是待估计的回归系数， ε 是随机误差

Model

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]
$R^2=0.957 \quad F=226 \quad p=0.000$		

结果解释

- 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
- a4置信区间包含零点，解释不可靠！

结果分析:残差分析方法



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

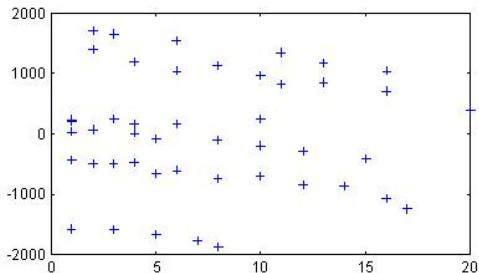


$$e = y - \hat{y}$$

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

e 与资历 x_1 的关系



映。

管理与教育的组合

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

进一步的模型

- 应在模型中增加管理 x_2 与教育 x_3, x_4 的交互项



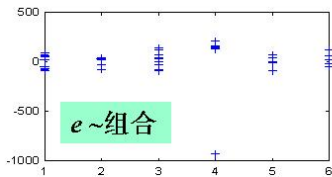
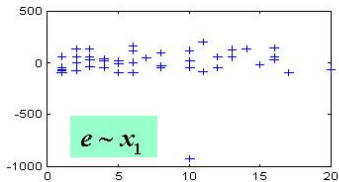
$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_2x_3 + a_6x_2x_4 + \varepsilon$$

- 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

Model

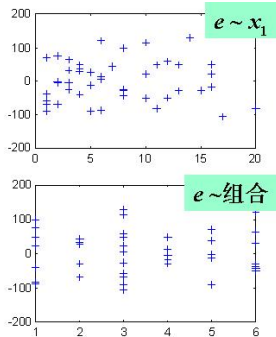
参数	参数估计值	置信区间
a_0	11204	[11044 11363]
a_1	497	[486 508]
a_2	7048	[6841 7255]
a_3	-1727	[-1939 -1514]
a_4	-348	[-545 -152]
a_5	-3071	[-3372 -2769]
a_6	1836	[1571 2101]
$R^2=0.999$ $F=554$ $p=0.000$		

Model



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]
$R^2= 0.9998 \quad F=36701 \quad p=0.0000$		



去掉异常数据后的结果

- 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

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

酶促反应

- 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
- 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：

Model

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	/

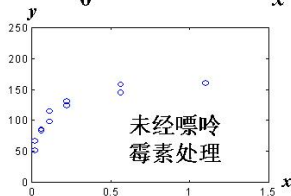
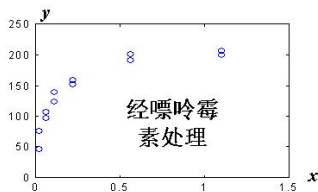
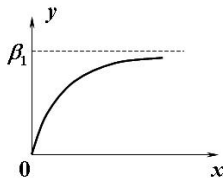
酶促反应的基本性质

- 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
-

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

Model

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



线性化模型

•

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

• 对 β_1, β_2 非线性, 对 θ_1, θ_2 线性

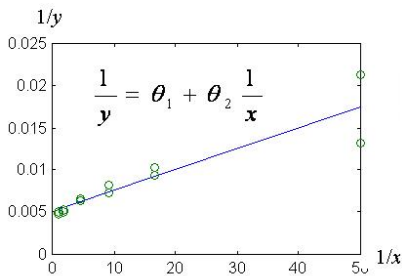
计算结果

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	5.107	[3.539 6.676]
θ_2	0.247	[0.176 0.319]
$R^2=0.8557$ $F=59.2975$ $p=0.0000$		

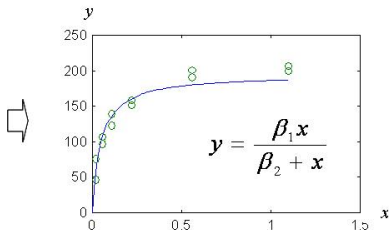
$$\hat{\beta}_1 = 1/\hat{\theta}_1 = 195.8027 \quad \hat{\beta}_2 = \hat{\theta}_2/\hat{\theta}_1 = 0.04841$$

计算结果分析



$1/x$ 较小时有很好的
线性趋势， $1/x$ 较大
时出现很大的起落

计算结果分析



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

- $[beta,R,J] = nlinfit(x,y,'model',beta0)$
- x , 自变量数据矩阵
- y , 因变量数据向量
- $model$, 模型的函数M文件名
- $beta0$, 给定的参数初值

非线性模型参数估计

- β , 参数的估计值; R , 残差; J , 估计预测误差的Jacobi矩阵
- β 的置信区间
- $\text{betaci} = \text{nlparci}(\beta, R, J)$

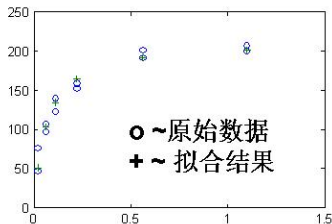
Matlab代码

- 先写好函数: `function y=f1(beta, x)`
- `y=beta(1)*x./(beta(2)+x);`
- 另建新的M文件: `x= ; y= ;beta0=[195.8027 0.04841];`
- `[beta,R,J]=nlinfit(x,y,' f1' ,beta0);`
- `betaci=nlparci(beta,R,J);`
- `nlintool(x,y,'f1',beta);`

非线性模型结果分析

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 $\hat{\beta}_2 = 0.0641$



投资额模型

- 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

该地区连续20年的统计数据

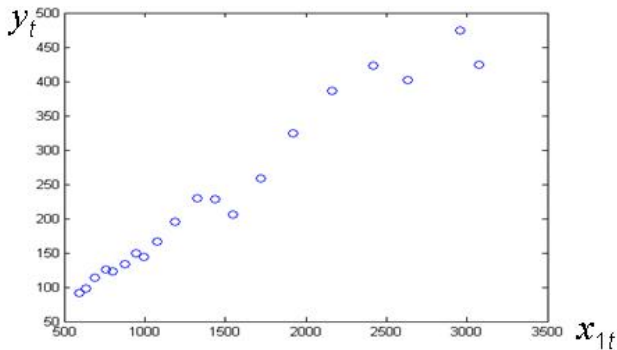
年份 序号	投资额	国民生产 总值	物价 指数	年份 序号	投资额	国民生 产总值	物价 指数
1	90.9	596.7	0.7167	11	229.8	1326.4	1.0575
2	97.4	637.7	0.7277	12	228.7	1434.2	1.1508
3	113.5	691.1	0.7436	13	206.1	1549.2	1.2579
4	125.7	756.0	0.7676	14	257.9	1718.0	1.3234
5	122.8	799.0	0.7906	15	324.1	1918.3	1.4005
6	133.3	873.4	0.8254	16	386.6	2163.9	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

基本回归模型

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

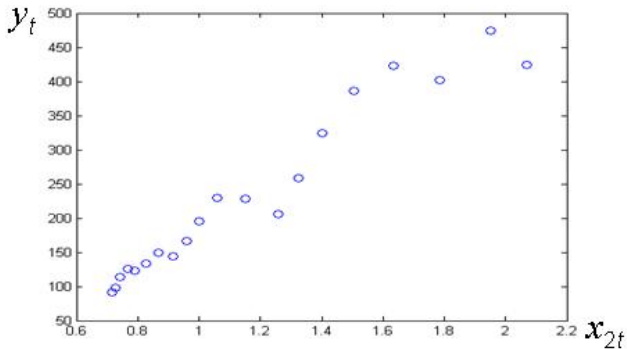
- 1 牙膏的销售量
- 2 软件开发人员的薪金
- 3 酶促反应
- 投资额模型

观察散点图



- 1 牙膏的销售量
- 2 软件开发人员的薪金
- 3 酶促反应
- 投资额模型

观察散点图



基本回归模型

- 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
β_0	322.7250	[224.3386 42
β_1	0.6185	[0.4773 0.
β_2	-859.4790	[-1121.4757 -5
$R^2= 0.9908 \quad F= 919.8529 \quad p=0.00$		

基本回归模型的结果与分析



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

- 剩余标准差 $s=12.7164$
- 模型优点: $R^2 = 0.9908$, 拟合度高
- 试考虑这样一个问题: 如果将投资的数据打乱, 我们前面模型结果是否还是一样?
- 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;
- 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

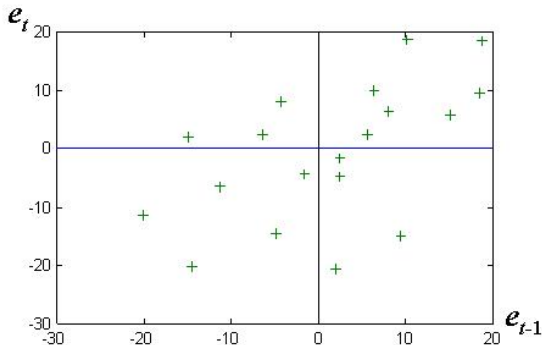
自相关性的定性诊断:残差诊断法

- 模型残差

$$e_t = y_t - \hat{y}_t$$

- 大部分点落在第1, 3象限: e_t 存在正的自相关
- 大部分点落在第2, 4象限: e_t 存在负的自相关

直观判断：基本回归模型的随机误差项 e_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, $|\rho| \leq 1$
- 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; $\rho > 0$, 正自相关性; $\rho < 0$, 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。

广义差分变换



$$DW = 21 - \hat{\rho} \Rightarrow \hat{\rho} = 1 - \frac{DW}{2}$$

- 原模型:

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

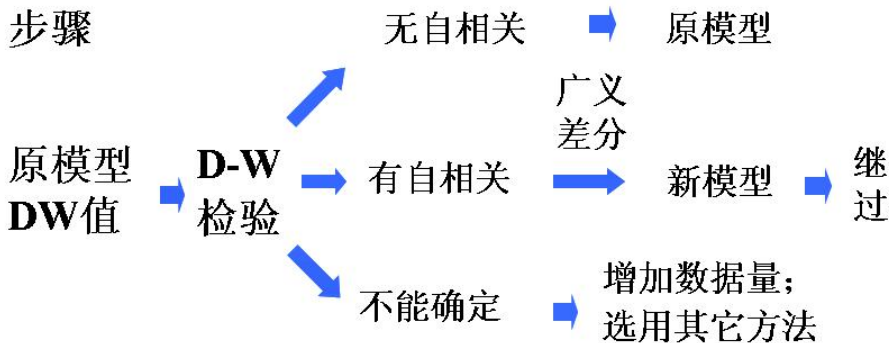
- 变换:

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

- 新模型是普通回归模型:

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

步骤



h投资额新模型的建立

- DW=0.8754, 临界值 (通过查表) $dL=1.10$, $dU=1.54$
- 原模型有正自相关
-

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

- 做变换:

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

-

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

新模型结果

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
β_0^*	163.4905	[1265.4592 2005.2178]
β_1	0.6990	[0.5751 0.8247]
β_2	-1009.0333	[-1235.9392 -782.1274]
$R^2= 0.9772 \quad F=342.8988 \quad p=0.0000$		

总体效果良好

剩余标准差

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

最终模型

- 经过检验，新模型无自相关性

-

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

- 还原变量:

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

- 一阶自回归模型

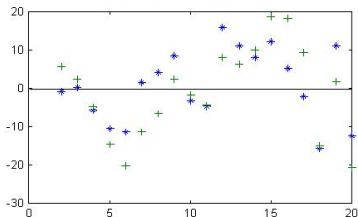
结果比较

模型结果比较

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

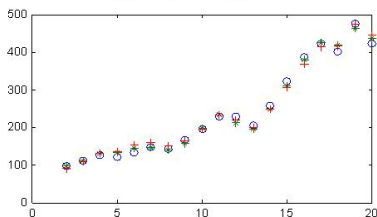
一阶自回归模型 $\hat{y}_t = 163.4905 + 0.5623y_{t-1} + 0.699x_{1,t} - 0.3930x_{1,t-1} - 1009.0333x_{2,t} + 567.3794x_{2,t-1}$

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 $\hat{y}_t \sim +$