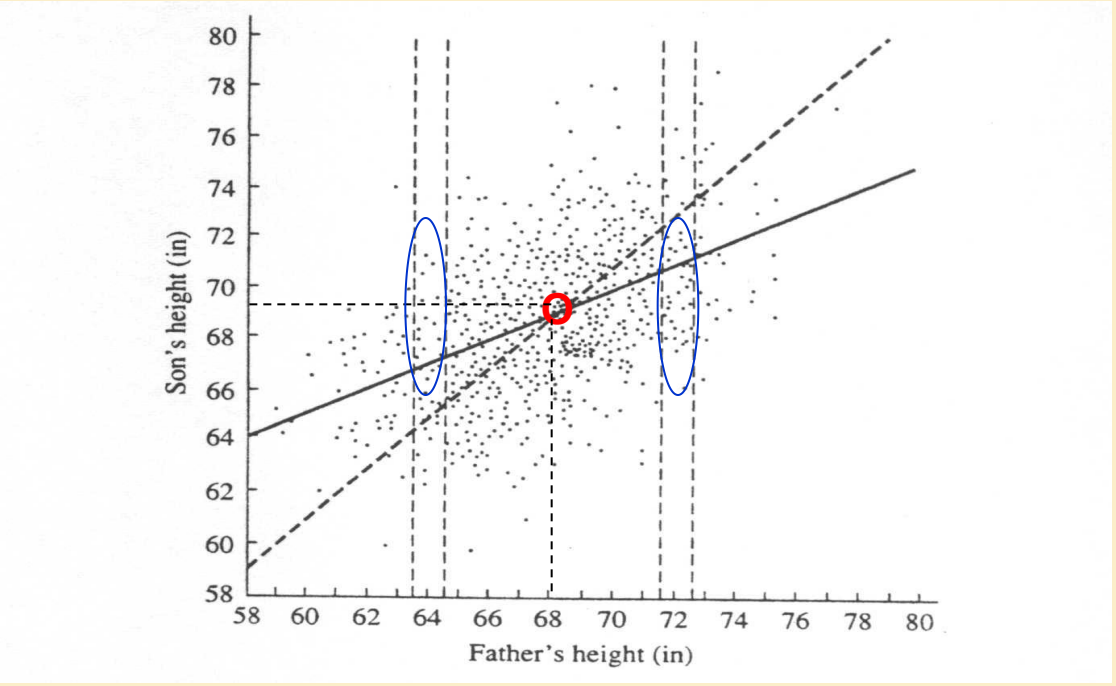
回归分析

回归（regression）？ Francis Golton (1822-1911)

- 一般说来高个子的父代会有高个子的子代
- 子代的身高比父代更加趋向一致(“向平庸的回归”)



$\bar{x} \approx 68, \bar{y} \approx 69$

儿子比父亲平均高1英寸

对于身高72英寸的父亲，
儿子身高多数不到73英寸;

对于身高64英寸的父亲，
儿子身高多数超过65英寸;

回归直线 $y=0.516\,x+33.73$

Pearson: 1078个父亲和儿子身高的散点图

回归分析是数学建模的有力工具

- 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；
- 人们关心的变量(因变量)受另外几个变量(自变量)的关联性(非因果性)的影响，并且存在众多随机因素，难以用机理分析方法找出它们之间的关系；
- 需要建立这些变量的数学模型，使得能够根据自变量的数值预测因变量的大小，或者解释因变量的变化。

血压与年龄　　刹车距离与车速

薪金与资历、教育程度、工作岗位

回归分析的主要步骤

- 收集一组包含因变量和自变量的数据；
- 选定因变量与自变量之间的模型，利用数据按照最小二乘准则计算模型中的系数；
- 利用统计分析方法对不同的模型进行比较，找出与数据拟合得最好的模型；
- 判断得到的模型是否适合于这组数据，诊断有无不适合回归模型的异常数据；
- 利用模型对因变量作出预测或解释。

回归分析(Regression Analysis)

- 从应用角度介绍回归分析的基本原理、方法和软件实现

1. 简化的实际问题及其数学模型
2. 一元线性回归
3. 多元线性回归
4. 非线性回归

实例及其数学模型　例1　血压与年龄

为了解血压随年龄增长而升高的关系，调查了30个成年人的血压（收缩压，mmHg）与年龄:

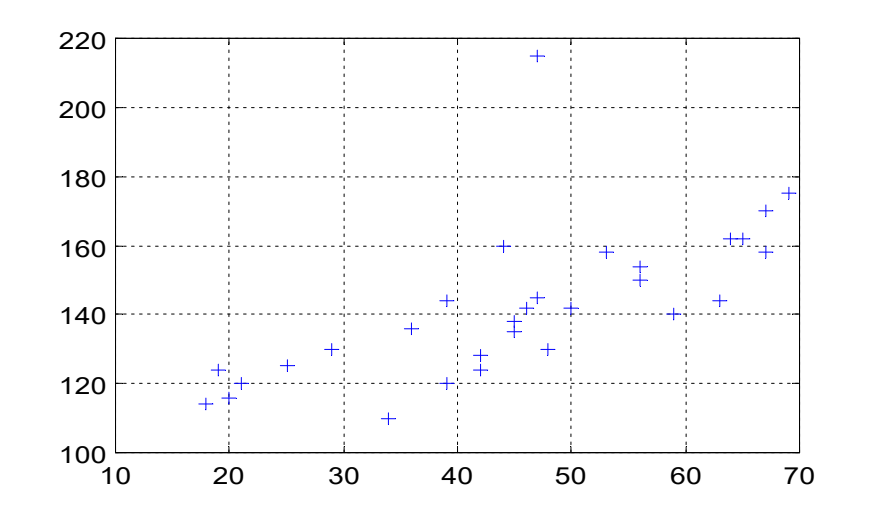
序号	血压	年龄	序号	血压	年龄	序号	血压	年龄
1	144	39	11	162	64	21	136	36
2	215	47	12	150	56	22	142	50
3	138	45	13	140	59	23	120	39
4	145	47	14	110	34	24	120	21
5	162	65	15	128	42	25	160	44
…	…	…	…	…	…	…	…	…

- 用这组数据确定血压与年龄的关系；
- 从年龄预测血压可能的变化范围；
- 回答“平均说来60岁比50岁的人血压高多少”。

例1 血压与年龄

模型 记血压(因变量) y，年龄(自变量) x，

作数据 $(x_i\ y_i)(i=1,2,\ldots 30)$的散点图



y与x大致呈线性关系

$$y=\beta_0+\beta_1 x$$

由数据确定系数 β_0,β_1

的估计值 $\hat{\beta}_0,\ \hat{\beta}_1$

- 曲线拟合(求超定线性方程组的最小二乘解);
- 从统计推断角度讨论β_0,β_1 的置信区间和假设检验;
- 对任意的年龄 x 给出血压 y 的预测区间。

例2 血压与年龄、体重指数、吸烟习惯

又调查了例1中30个成年人的体重指数、吸烟习惯:

序号	血压	年龄	体重指数	吸烟	序号	血压	年龄	体重指数	吸烟	序号	血压	年龄	体重指数	吸烟
1	144	39	24.2	0	11	162	64	28.0	1	21	136	36	25.0	0
2	215	47	31.1	1	12	150	56	25.8	0	22	142	50	26.2	1
3	138	45	22.6	0	13	140	59	27.3	0	23	120	39	23.5	0
4	145	47	24.0	1	14	110	34	20.1	0	24	120	21	20.3	0
5	162	65	25.9	1	15	128	42	21.7	0	25	160	44	27.1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

体重指数: 体重(kg) /[身高(m)]^2

吸烟习惯: 0~不吸烟，1~吸烟

例2 血压与年龄、体重指数、吸烟习惯

模型 记血压 y，年龄x_1、体重指数x_2、吸烟习惯x_3

作数据 y 对x_2 的散点图　　y与x_2大致呈线性关系

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

由数据确定系数 $\beta_0, \beta_1, \beta_2, \beta_3$

的估计值 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

例3　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
05	11767	1	0	3	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=研究生

模型　　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=研究生

$$x_3=\begin{cases}1, & \text{中学}\\ 0, & \text{其它}\end{cases}\qquad x_4=\begin{cases}1, & \text{大学}\\ 0, & \text{其它}\end{cases}$$

⇨ 中学：$x_3=1, x_4=0$；大学：$x_3=0, x_4=1$；研究生：$x_3=0, x_4=0$

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　$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$

由数据确定 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

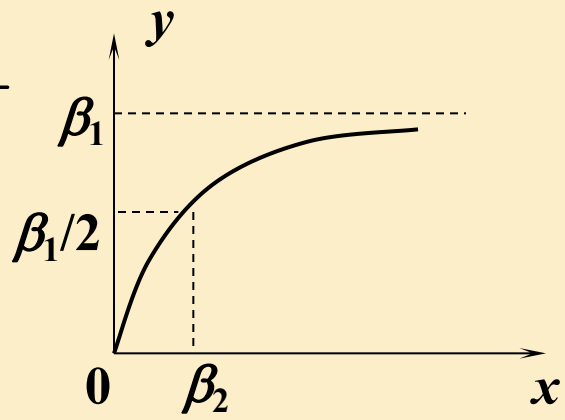
例4　酶促反应

酶~高效生物催化剂; 酶促反应~经过酶催化的化学反应

酶促反应的反应速度主要取决于反应物（底物）的浓度:

- 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.

Michaelis-Menten模型　$y = \dfrac{\beta_1 x}{\beta_2 + x}$



y ~ 酶促反应的速度, x ~底物浓度

待定系数 β_1 (最终反应速度)　　β_2 (半速度点)

例4　酶促反应

为研究酶促反应中嘌呤霉素对反应速度与底物浓度之间关系的影响，设计了两个实验：使用的酶经过嘌呤霉素处理；使用的酶未经嘌呤霉素处理。

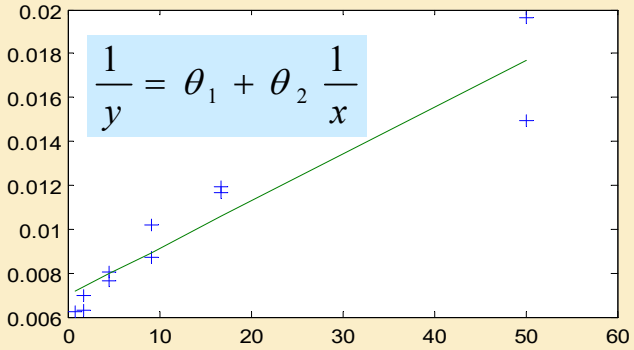
实验数据

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	/

对未经嘌呤霉素处理的反应，用实验数据估计参数β_1, β_2；用实验数据研究嘌呤霉素处理对参数β_1, β_2的影响。

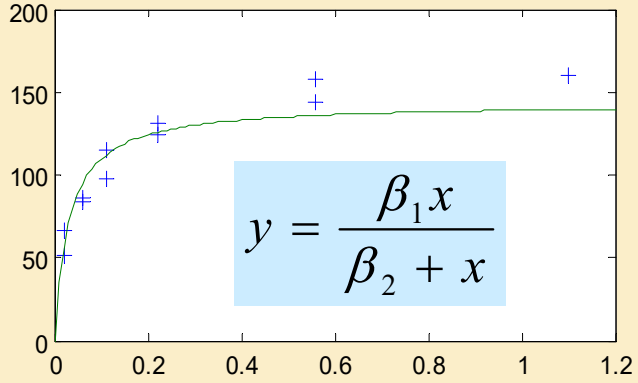
模型 $y=\dfrac{\beta_1 x}{\beta_2+x}$ ⇨ $\dfrac{1}{y}=\dfrac{1}{\beta_1}+\dfrac{\beta_2}{\beta_1}\dfrac{1}{x}=\theta_1+\theta_2\dfrac{1}{x}$

对β_1, β_2非线性　　对θ_1, θ_2线性



1/*x*较小时有很好的线性趋势，
1/*x*较大时出现很大的分散.

$\theta_1=6.972\times10^{-3}$, $\theta_2=0.215\times10^{-3}$ ⇨ $\beta_1=143.43$, $\beta_2=0.0308$



***x*较大时，*y*有较大偏差.**
参数估计时，*x*较小（1/*x*很大）的数据控制了参数的确定.

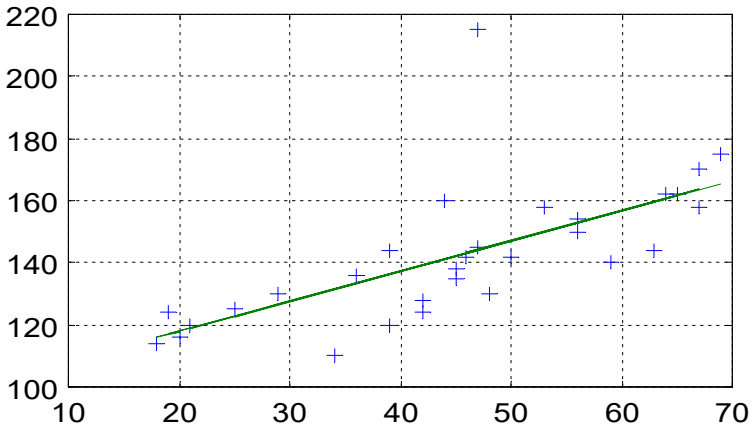
直接考虑非线性模型

一元线性回归分析

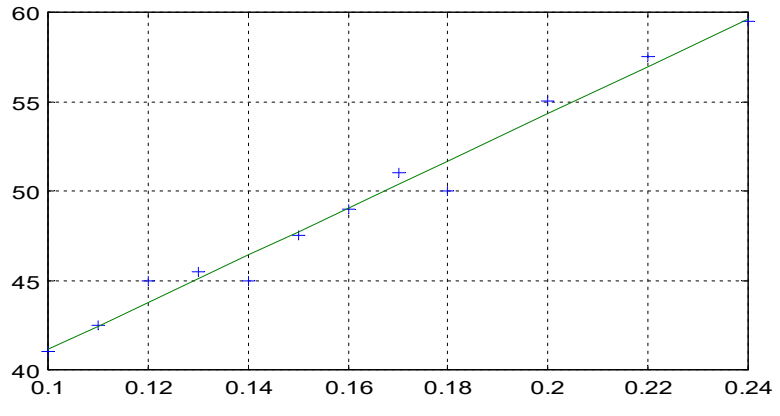
问题 已知一组数据 (x_i, y_i), $i=1,2,\ldots n$ (平面上的n个点)，

用最小二乘准则确定一个线性函数(直线) $y = \beta_0 + \beta_1 x$

1. 血压与年龄



2. 合金强度与碳含量



系数的计算二者没有什么区别；2的拟合效果比1好得多.

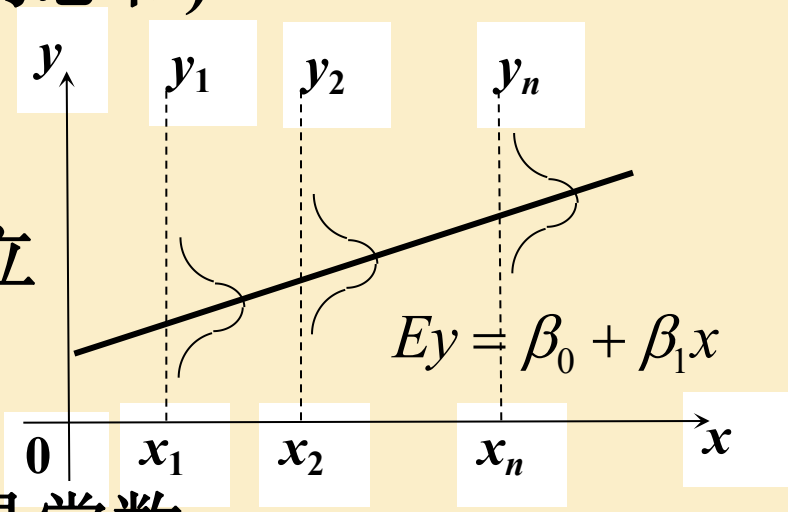
怎样衡量由最小二乘准则拟合得到的模型的可靠程度？

怎样给出模型系数的置信区间和因变量的预测区间？

一元线性回归模型 $y = \beta_0 + \beta_1 x + \varepsilon$

x~自变量　　β_0，β_1 ~回归系数

ε~随机变量(影响y的随机因素的总和)



基本假设

独立性: 对于不同的x，y相互独立

线性性: y的期望是x的线性函数

齐次性: 对于不同的x，y的方差是常数

正态性: 对于给定的x，y服从正态分布

ε是相互独立的、期望为0、方差为σ^2、正态分布的随机变量，即$\varepsilon \sim N(0, \sigma^2)$, ε称（随机）误差。

回归系数的最小二乘估计

数据x_i, y_i(i=1,…n)代入 $y=\beta_0+\beta_1 x+\varepsilon \Rightarrow y_i=\beta_0+\beta_1 x_i+\varepsilon_i$

误差平方和 $Q(\beta_0,\beta_1)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2$

$$\frac{\partial Q}{\partial \beta_0}=0,\ \frac{\partial Q}{\partial \beta_1}=0 \quad \Rightarrow \quad \hat{\beta}_1=\frac{s_{xy}}{s_{xx}},\quad \hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$$

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\ \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,\ s_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2,\ s_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

直线 $y=\hat{\beta}_0+\hat{\beta}_1 x$ 通过x_i, y_i的均值点 $(\bar{x},\bar{y})$

最小二乘估计　　线性无偏最小方差估计

一元线性回归的统计分析

1.误差方差$D\varepsilon=\sigma^2$ 的估计

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i$$

$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i, i=1,2,\cdots n$　　y_i理论值(期望)的估计

$\hat{\varepsilon}_i=y_i-\hat{y}_i, i=1,2,\cdots n$　　误差ε_i的估计,称残差(记作e_i)

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

σ^2的无偏估计　　$s^2=\hat{\sigma}^2=\dfrac{Q}{n-2}$

n-2~Q的自由度=数据容量 – 模型中所含参数的个数

s^2~剩余方差(样本方差)，s~剩余标准差(样本标准差)

一元线性回归的统计分析

2. 回归系数的区间估计和假设检验

统计性质：$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / s_{xx})$，$Q/\sigma^2 \sim \chi^2_{(n-2)}$，$\hat{\beta}_1$和$Q$相互独立

t 分布　$$t = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{s_{xx}}/\sigma}{\sqrt{Q/(n-2)\sigma^2}} = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{s_{xx}}}{s} \sim t_{(n-2)}$$

β_1的置信区间　$$[\hat{\beta}_1 - t_{(n-2),1-\alpha/2}\frac{s}{\sqrt{s_{xx}}},\ \hat{\beta}_1 + t_{(n-2),1-\alpha/2}\frac{s}{\sqrt{s_{xx}}}]$$

问：怎样缩短β_1的置信区间？

对β_1的假设检验　$H_0 : \beta_1 = 0,\quad H_1 : \beta_1 \neq 0$

$$|t| = \left|\frac{\hat{\beta}_1\sqrt{s_{xx}}}{s}\right| > t_{(n-2),1-\alpha/2} \Rightarrow \text{拒绝}H_0 \Rightarrow \text{回归模型有效} \Leftrightarrow \beta_1\text{的置信区间不包含零点}$$

一元线性回归的统计分析

3.模型的有效性检验　　偏差的分解：$y_i-\bar{y}=(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})$

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{S}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{Q}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{U}$$

S = Q + U

总偏差平方和　　残差平方和　　回归平方和

决定系数 $R^2=U/S$　　因变量的总变化中自变量引起的部分的比例

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i \qquad \bar{y}=\hat{\beta}_0+\hat{\beta}_1\bar{x} \qquad U=\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2=\hat{\beta}_1^2s_{xx}$$

若H_0成立 $U/\sigma^2=\hat{\beta}_1^2s_{xx}/\sigma^2\sim\chi^2_{(1)}$

$$Q/\sigma^2\sim\chi^2_{(n-2)},$$

$$\Rightarrow F=\frac{U}{Q/(n-2)}\sim F_{(1,n-2)}$$

给定α，有$F_{(1,n-2),\,1-\alpha}$　　$F>F_{(1,n-2),\,1-\alpha}$ ⇒ 拒绝H_0 ⇒ 回归模型有效

利用一元线性回归模型进行预测

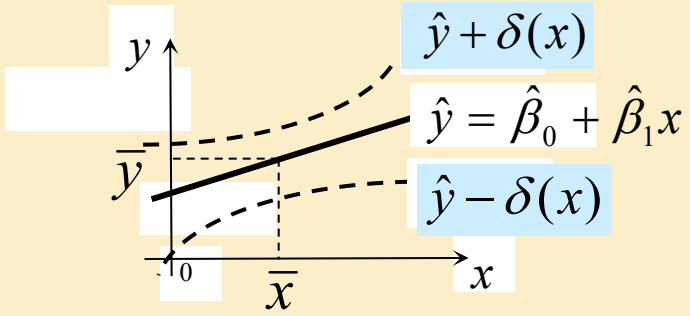
x_0给定，y_0的预测值：　$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

性质：$\hat{y}_0$ 无偏，且 $E(\hat{y}_0 - y_0)^2$ 最小

预测区间

$$[\hat{y}_0 - t_{(n-2),1-\alpha/2} s \sqrt{\frac{(x_0-\bar{x})^2}{s_{xx}} + \frac{1}{n} + 1},\ \hat{y}_0 + t_{(n-2),1-\alpha/2} s \sqrt{\frac{(x_0-\bar{x})^2}{s_{xx}} + \frac{1}{n} + 1}]$$

s~剩余标准差



n很大且 x_0接近 $\bar{x}$

$$[\hat{y}_0 - u_{1-\alpha/2} s,\ \hat{y}_0 + u_{1-\alpha/2} s]$$

$$\delta(x) = t_{(n-2),1-\alpha/2} s \sqrt{\frac{(x-\bar{x})^2}{s_{xx}} + \frac{1}{n} + 1} \approx u_{1-\alpha/2} s$$

一元线性回归的MATLAB实现

b=regress(y,X)

[b,bint,r,rint,s]=regress(y,X,alpha)

输入：**y~**因变量（列向量），**X~1**与自变量组成的矩阵，**alpha~**显著性水平α（缺省时设定为**0.05**）。

输出：$b=(\hat{\beta}_0,\hat{\beta}_1)$，**bint~**$\beta_0$, β_1的置信区间，**r~**残差（列向量），**rint~**残差的置信区间，

s(3个统计量和误差方差的估计)：决定系数R^2; F值; $F_{(1,n-2)}$分布的分位数 $F_{(1,n-2),\,1-\alpha}$ 大于F值的概率p。当 $p<\alpha$ 时拒绝H_0，回归模型有效。

注意 **regress** 与 **polyfit** 用法的区别

例1 血压与年龄 模型 $y = \beta_0 + \beta_1 x$ 数据

xueya1.m

回归系数	回归系数估计值	回归系数置信区间
β_0	98.4084	[78.7484 118.0683]
β_1	0.9732	[0.5601 1.3864]
R^2=0.4540 F=23.2834 p<0.0001 s^2 = 273.7137		

模型检验

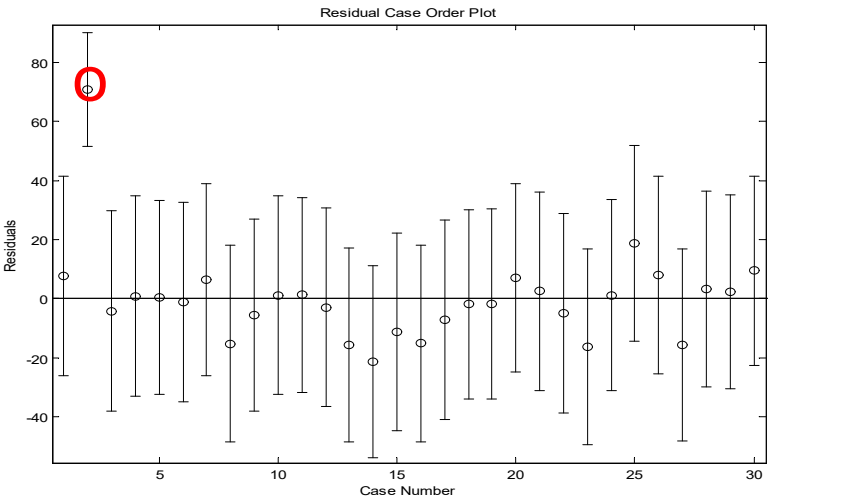
β_1置信区间不含零点；$p<\alpha$；$F_{(1,n-2),\,1-\alpha} = 4.1960 < F$

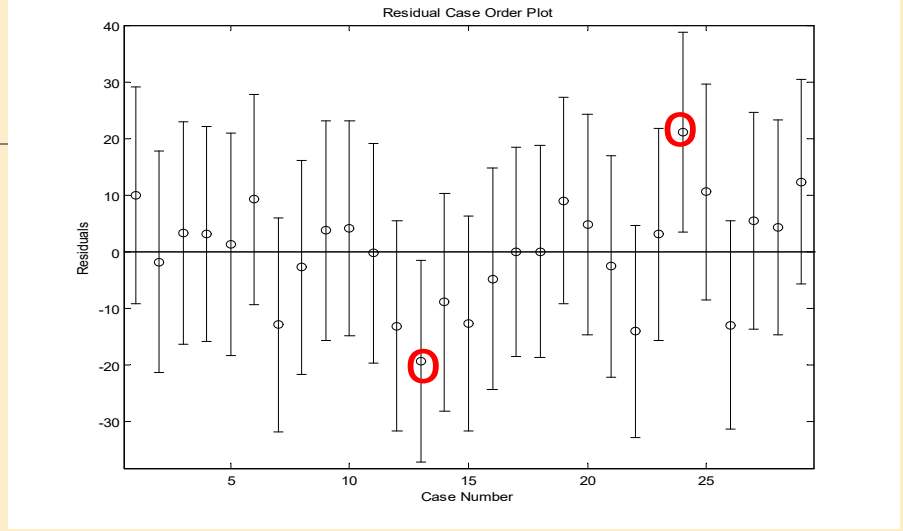
β_1置信区间较长，R^2较小，模型精度不高。

由残差图剔除异常数据后

回归系数	回归系数估计值	回归系数置信区间
β_0	96.8665	[85.4771 108.2559]
β_1	0.9533	[0.7140 1.1925]
R^2= 0.7123 F= 66.8358 p<0.0001 s^2 =91.4305		

例1 血压与年龄 模型 $y = \beta_0 + \beta_1 x$ xueya.m





剔除异常点 (x_2, y_2) 又出现两个新的异常点.

对50岁人的血压进行预测： $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 144.5298$

预测区间 (α=0.05)： [124.5406 164.5190]

简化 (t→u)： [125.7887 163.2708]

多元线性回归分析

模型 $y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$

估计回归系数 ⇩ $(y_i,x_{i1},\cdots x_{im}),\ i=1,\cdots n,\ n>m$

$$y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_mx_{im}+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma^2),\ i=1,\cdots n$$

⇩

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

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

$$Q(\beta)=\sum_{i=1}^{n}\varepsilon_i^2=(Y-X\beta)^T(Y-X\beta)\qquad \frac{\partial Q}{\partial\beta_i}=0,\ i=0,1,\cdots m$$

⇩ $X^T(Y-X\beta)=0$ ⇩ $\hat{\beta}=(X^TX)^{-1}X^TY$ 最小二乘估计

思考 怎样保证 X^TX 可逆 为什么要求 $n>m$

多元线性回归的统计分析

1.误差方差σ^2的估计

	一元回归	多元回归
模型	$y=\beta_0+\beta_1 x_1+\varepsilon$	$y=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m+\varepsilon$
估计值	$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$	$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\cdots+\hat{\beta}_m x_{mi}$
残差	$e_i=\hat{\varepsilon}_i=y_i-\hat{y}_i$	$e_i=\hat{\varepsilon}_i=y_i-\hat{y}_i$
残差平方和	$Q=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$Q=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
剩余方差	$s^2=\hat{\sigma}^2=\frac{Q}{n-2}$	$s^2=\hat{\sigma}^2=\frac{Q}{n-m-1}$
Q的自由度	n-2 (2个参数)	n-(m+1) (m+1个参数)

2. 回归系数的区间估计和假设检验

一元回归

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / s_{xx}), \quad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$Q / \sigma^2 \sim \chi^2_{(n-2)},$$

$$t = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{s_{xx}}/\sigma}{\sqrt{Q/(n-2)\sigma^2}} = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{s_{xx}}}{s} \sim t_{(n-2)}$$

$$[\hat{\beta}_1 - t_{(n-2),1-\alpha/2}\frac{s}{\sqrt{s_{xx}}}, \ \hat{\beta}_1 + t_{(n-2),1-\alpha/2}\frac{s}{\sqrt{s_{xx}}}]$$

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$$

$$|t| = \left|\frac{\hat{\beta}_1\sqrt{s_{xx}}}{s}\right| > t_{(n-2),1-\alpha/2}$$

多元回归

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj}), \quad c_{jj} \sim (\tilde{X}^T \tilde{X})^{-1}$$

的j对角元

$$Q / \sigma^2 \sim \chi^2_{(n-m-1)}$$

$$t_j = \frac{(\hat{\beta}_j - \beta_j)/\sigma\sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)\sigma^2}} = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim t_{(n-2)}$$

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

$$H_0^{(j)} : \beta_j = 0, \quad H_1^{(j)} : \beta_j \neq 0$$

$$|t_j| = \left|\frac{\hat{\beta}_j}{s\sqrt{c_{jj}}}\right| > t_{(n-2),1-\alpha/2}$$

拒绝H_0, 模型有效

3. 模型的有效性检验

	一元回归	多元回归
偏差分解	$S = U + Q$	$S = U + Q$
决定系数	$R^2 = U/S$	$R^2 = U/S$
	$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$	$H_0^{(j)} : \beta_j = 0, \quad H_1^{(j)} : \beta_j \neq 0$
H_0成立	$U/\sigma^2 \sim \chi^2_{(1)}, \quad Q/\sigma^2 \sim \chi^2_{(n-2)},$	$U/\sigma^2 \sim \chi^2_{(m)},$
	$F = \dfrac{U}{Q/(n-2)} \sim F_{(1,n-2)}$	$F = \dfrac{U/m}{Q/(n-m-1)} \sim F_{(m,n-m-1)}$
检验	$F > F_{(1,n-2),\,1-\alpha}$	$F_{(m,\,n-m-1),\,1-\alpha}$

$Q/\sigma^2 \sim \chi^2_{(n-m-1)}$

拒绝H_0, 模型有效

利用多元线性回归模型进行预测

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$$

性质：$\hat{y}_0$ 无偏，且 $E(\hat{y}_0 - y_0)^2$ 最小

预测区间 $[\hat{y} - \delta(x),\ \hat{y} + \delta(x)]$

$$\delta(x) = t_{(n-2),1-\alpha/2} s \sqrt{\frac{(x-\bar{x})^2}{s_{xx}} + \frac{1}{n} + 1} \approx u_{1-\alpha/2} s$$

$$\delta(x) = t_{(n-2),1-\alpha/2} s \sqrt{(x-\bar{x})^T (\tilde{X}^T \tilde{X})^{-1} (x-\bar{x}) + \frac{1}{n} + 1} \approx u_{1-\alpha/2} s$$

与一元回归对比

多元线性回归的MATLAB实现

与一元回归相同

b=regress(y,X)　　注意 X 的构造

[b,bint,r,rint,s]=regress(y,X,alpha)

例2 血压与年龄、体重指数、吸烟习惯　xueya2.m

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$　　剔除两个异常点后

$$\hat{y} = 58.5101 + 0.4303x_1 + 2.3449x_2 + 10.3065x_3$$

- 年龄和体重指数相同，吸烟者比不吸烟者的血压(平均)高10. 3
- 与例1“血压与年龄”的结果 $\hat{y} = 96.8665 + 0.9533x_1$ 相比，

  年龄增加1岁血压的升高值(即 β_1)为何有这么大的差别

线性最小二乘拟合与多元线性回归的一般形式

线性回归模型 $y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$ (1)

“线性”是指***y***是系数β的关系(非指***y***与***x***的关系)

$y=\beta_0+\beta_1x^2,\quad y=\beta_0+\beta_1e^{x_1}+\beta_2/x_2$ ~ 线性回归

线性回归一般形式 $y=\beta_0+\beta_1r_1(x)+\cdots+\beta_mr_m(x)+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$ (2)

$x=(x_1,\cdots x_k),\quad r_j(x)\ (j=1,\cdots,m)$ 是已知函数

令$r_j(x)=u_j$, 则(2)→(1)

多元线性回归中的交互作用

例3 软件开发人员的薪金 $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$

y~薪金, x_1~资历, $x_2=1$~ 管理人员, $x_2=0$~ 非管理人员

x_3=1, x_4=0~中学； x_3=0, x_4=1~大学； x_3=0, x_4=0~研究生

系数	系数估计	置信区间
β_0	11032	[10258 11807]
β_1	546	[484 608]
β_2	6883	[6248 7517]
β_3	-2994	[-3826 -2162]
β_4	148	[-636 931]
R^2=0.957	F=226	p=0.000

xinjin1.m

资历增加1年
薪金增长546

管理人员多6883

中学程度比更高的少2994

大学程度比更高的多148

β_4置信区间包含零点，
解释不可靠！

$R^2, F, p\rightarrow$ 模型整体上可用

用残差分析发现交互作用

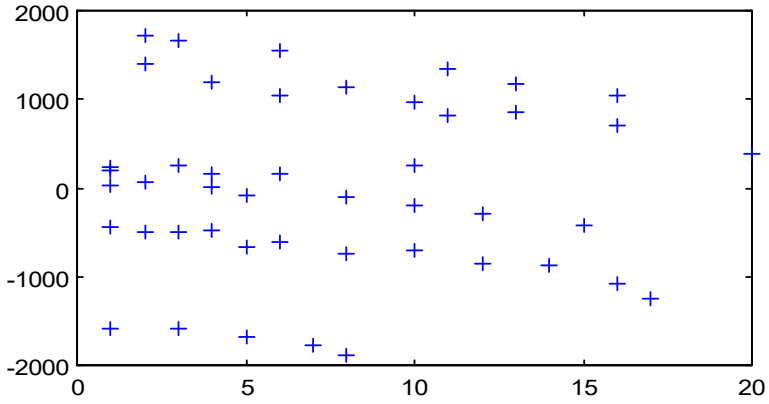
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

考察残差 $e = y - \hat{y}$ 是否为$N(0, \sigma^2)$

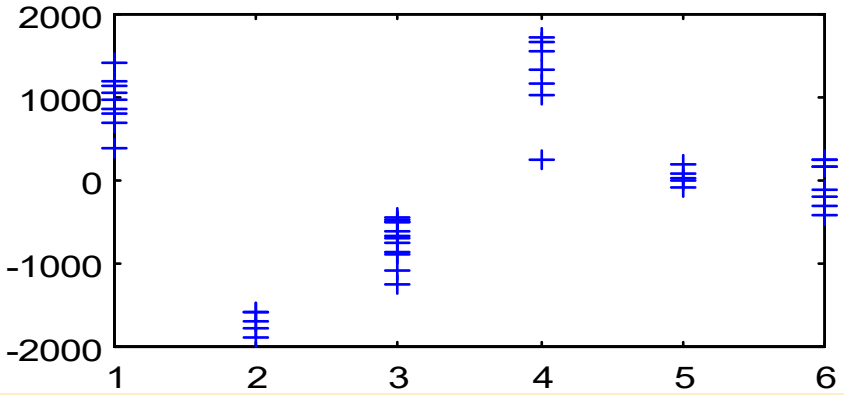
管理与教育的组合

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

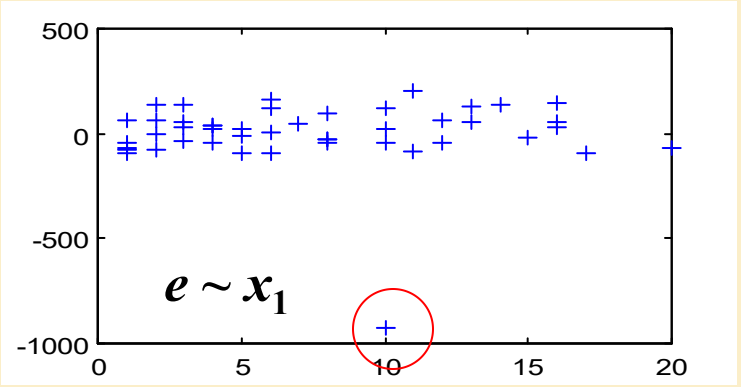
应增加x_2与x_3, x_4的交互项

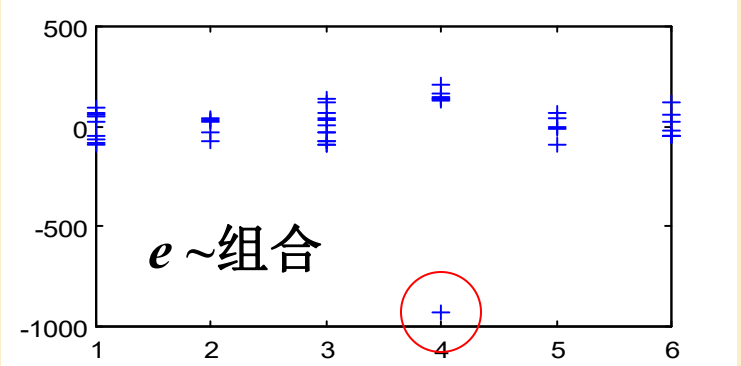
增加管理x_2与教育x_3, x_4的交互项

xinjin2.m

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_2 x_3 + \beta_6 x_2 x_4 + \varepsilon$$

系数	系数估计值	置信区间
β_0	11204	[11044 11363]
β_1	497	[486 508]
β_2	7048	[6841 7255]
β_3	-1727	[-1939 -1514]
β_4	-348	[-545 –152]
β_5	-3071	[-3372 -2769]
β_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

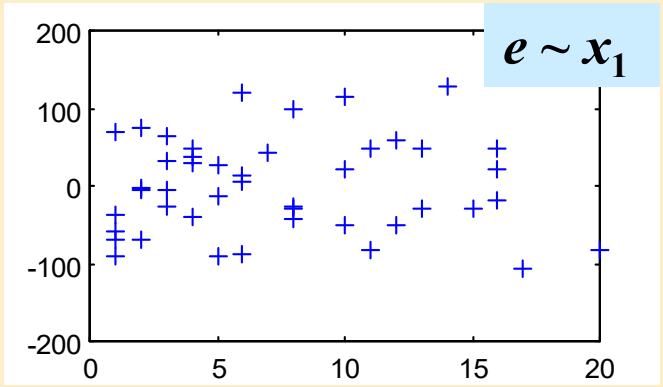
消除了不正常现象

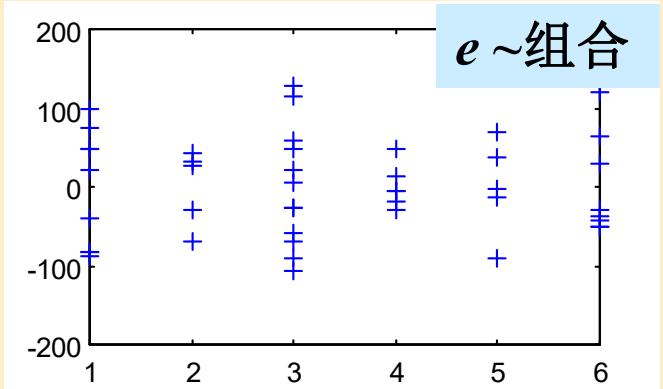
异常数据(33号)应去掉

去掉异常数据后的结果

系数	系数估计值	置信区间
β_0	11200	[11139 11261]
β_1	498	[494 503]
β_2	7041	[6962 7120]
β_3	-1737	[-1818 -1656]
β_4	-356	[-431 –281]
β_5	-3056	[-3171 –2942]
β_6	1997	[1894 2100]
R^2= 0.9998 F=36701 p=0.0000		

xinjin3.m





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

制订6种管理—教育组合人员的“基础”薪金(资历$x_1=0$)

组合	管理x_2	教育 (x_3, x_4)	系数	“基础”薪金
1	0	(1,0)	$\beta_0+\beta_3$	9463
2	1	(1,0)	$\beta_0+\beta_2+\beta_3+\beta_5$	13448
3	0	(0,1)	$\beta_0+\beta_4$	10844
4	1	(0,1)	$\beta_0+\beta_2+\beta_4+\beta_6$	19882
5	0	(0,0)	β_0	11200
6	1	(0,0)	$\beta_0+\beta_2$	18241

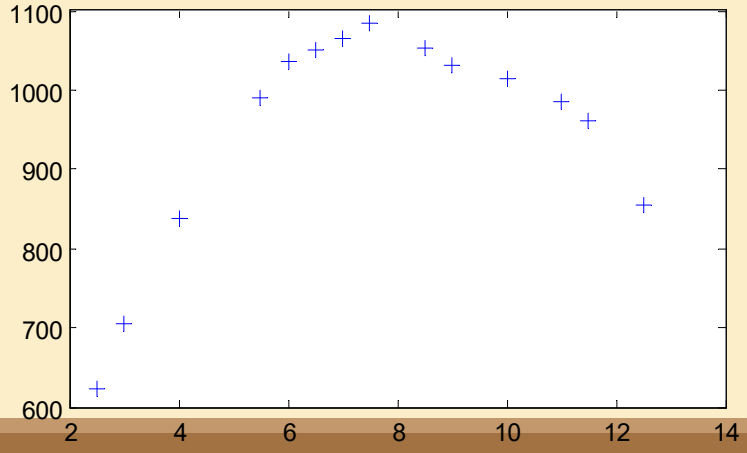
大学程度管理人员比更高程度管理人员的薪金高

大学程度非管理人员比更高程度非管理人员的薪金略低

线性回归的特殊情形----多项式回归

例1 西红柿的施肥量与产量 14块同样大小土地的数据

序号	产量(升)	施肥(千克)	序号	产量(升)	施肥(千克)
1	1035	6.0	…	…	…
2	624	2.5	12	1030	9.0
3	1084	7.5	13	985	11.0
…	…	…	14	855	12.5



模型 $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

b=regress(y,X)求解

$$\hat{y} = 175.62 + 217.87x - 13.15x^2$$

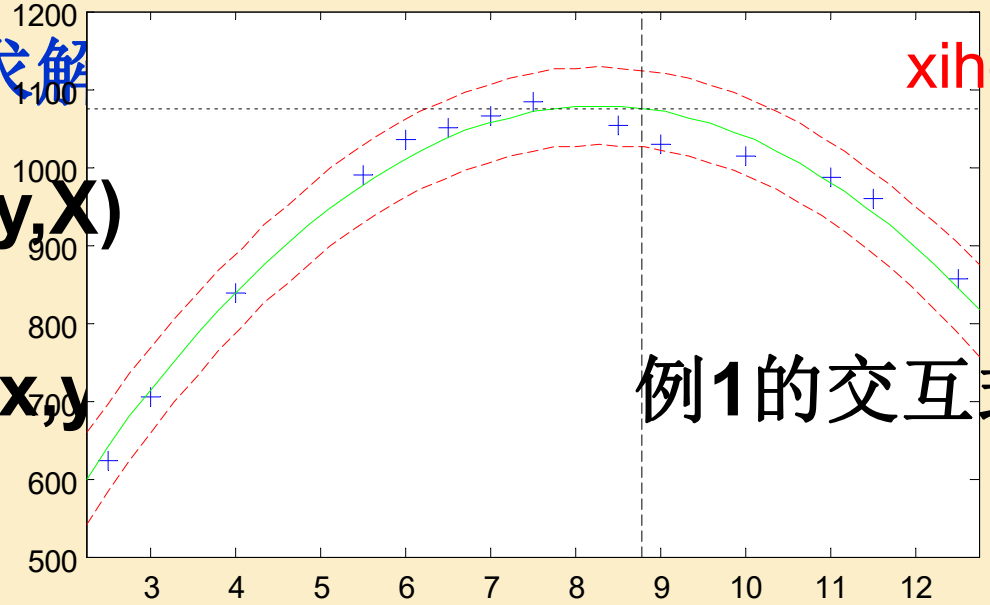
一元多项式回归模型的一般形式

$$y = \beta_0 + \beta_1 x + \cdots + \beta_m x^m + \varepsilon$$

MATLAB求解

xihongshi.m

- **regress(y,X)**
- **polytool(x,y**



例1的交互式画面

注意3个程序的用法与所得结果的相同点和不同点

例2　商品销售量与价格

某厂生产的一种电器的销售量y与竞争对手的价格x_1和本厂的价格x_2有关。

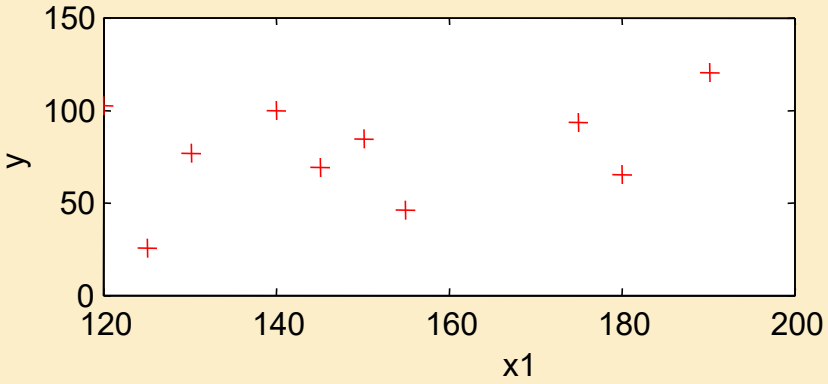
下表是该商品在10个城市的销售记录。

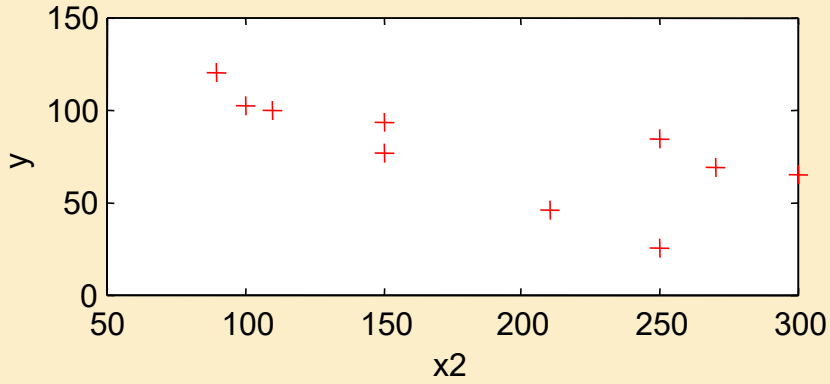
x_1（元）	120	140	190	130	155	175	125	145	180	150
x_2（元）	100	110	90	150	210	150	250	270	300	250
y（个）	102	100	120	77	46	93	26	69	65	85

1）根据这些数据建立 y 与 x_1和 x_2的关系式，对得到的模型和系数进行检验。

2）若某市本厂产品售价160元，竞争对手售价170元，预测该市的销售量。

例2　商品销售量与价格





y与x_2有较明显的线性关系，y与x_1的关系难以确定.

需要试验不同的回归模型，用统计分析决定优劣.

线性模型 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

shangpin.m

系数	系数估计	置信区间
β_0	66.5176	[-32.5060 165.5411]
β_1	0.4139	[-0.2018 1.0296]
β_2	-0.2698	[-0.4611 -0.0785]
R^2= 0.6527, F=6.5786, p= 0.0247,s^2= 351.0445		

置信区间包含零点

整体检验效果不好

二次函数回归模型

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

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

多元二项式回归的一般形式

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

MATLAB命令：rstool(x,y,'model',alpha)

X~ n×m自变量矩阵, y~因变量向量，model选择:

linear（只包含线性项)）；

purequadratic（包含线性项和纯二次项）；

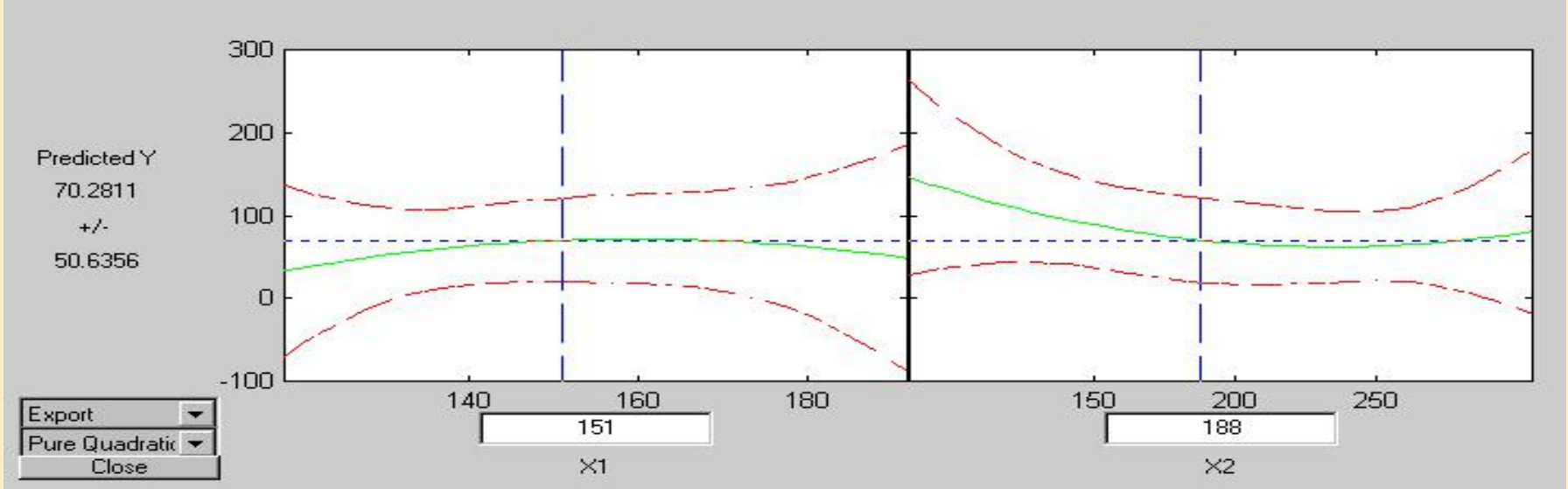
interaction（包含线性项和纯交互项）；

quadratic（包含线性项和完全二次项）。

输出一个交互式画面

例2　商品销售量与价格

model = purequadratic



4个模型的输出比较

	β_0	β_1	β_2	β_3	β_4	β_5	s
Purequadratic	-312.5871	7.2701	-1.7337	-0.0228	0.0037		16.6436
quadratic	-307.3600	7.2032	-1.7374	0.0001	-0.0226	0.0037	18.6064
interaction	137.5317	-0.0372	-0.7131	0.0028			19.1626
linear	66.5176	0.4139	-0.2698				18.7362

变量选择与逐步回归

变量选择　　影响因变量的因素：

自变量$x_1, x_2, \ldots x_m$及其简单函数, 如 $x_i^2, 1/x_i, e^{x_i}\ (i \in \{1,2,\cdots m\})$

- 将所有影响显著的因素都纳入回归模型；
- 最终的模型尽量简单, 即包含尽量少的因素。

变量选择的标准　　$s^2 = Q/(n-p-1),\ s^2$最小

- 从候选集合S=$\{x_1,\ldots x_k\}$中选出一子集S_1 (含$p \le k$个自变量)与因变量y构造回归模型, 其优劣由s^2度量.
- 影响显著的自变量进入模型时，Q明显下降，s减小；
- 影响很小的自变量进入模型时，Q下降不大，p的增加会使s变大.

逐步回归

- 从候选集合中确定一初始子集；
- 从子集外（候选集合内）中引入一个对y影响显著的；
- 对集合中的变量进行检验，剔除影响变得不显著的；
- 迭代式地进行引入和剔除，直到不能进行为止。
- 选择衡量影响显著程度的统计量，通常用偏F统计量；
- 适当选取引入变量的显著性水平α_{in}和剔除变量的α_{out}。
- 引入新的变量后原来模型内影响显著的变量变得不显著，从而被剔除～自变量之间存在较强相关性的结果.

多重共线性　　某些自变量之间的相关性很强

⇨矩阵X^TX病态　⇨回归系数的置信区间较大

MATLAB中的逐步回归

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

x~n×k自变量数据矩阵**(k~**全部变量数**), y~**因变量向量,

inmodel~初始模型中候选变量的指标（**x**的列序数，缺省时为全部候选变量），

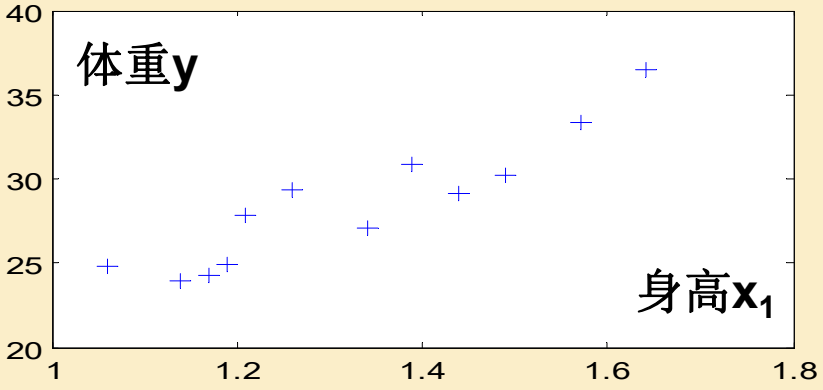
penter ~ 引入变量的显著性水平α_{in} **(**缺省时为**0.05)**

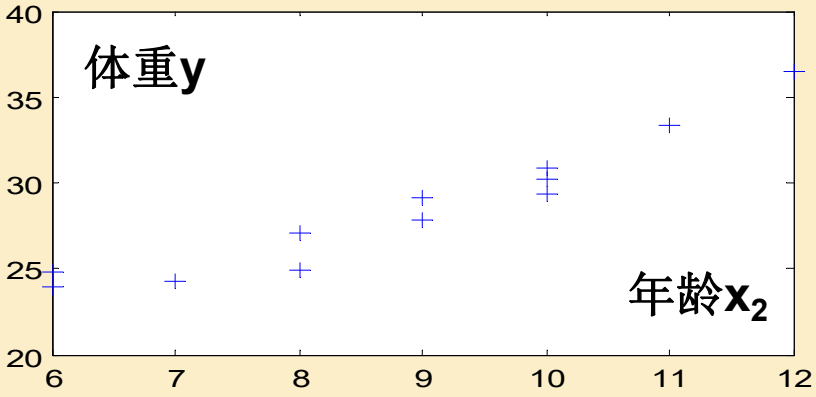
premove~剔除变量的显著性水平α_{out} **(**缺省时为**0.10)**

输出交互式画面

例 儿童的体重与身高和年龄

序号	体重(kg)	身高(m)	年龄	序号	体重(kg)	身高(m)	年龄
1	27.1	1.34	8	7	30.9	1.39	10
2	30.2	1.49	10	8	27.8	1.21	9
3	24.0	1.14	6	9	29.4	1.26	10
4	33.4	1.57	11	10	24.8	1.06	6
5	24.9	1.19	8	11	36.5	1.64	12
6	24.3	1.17	7	12	29.1	1.44	9



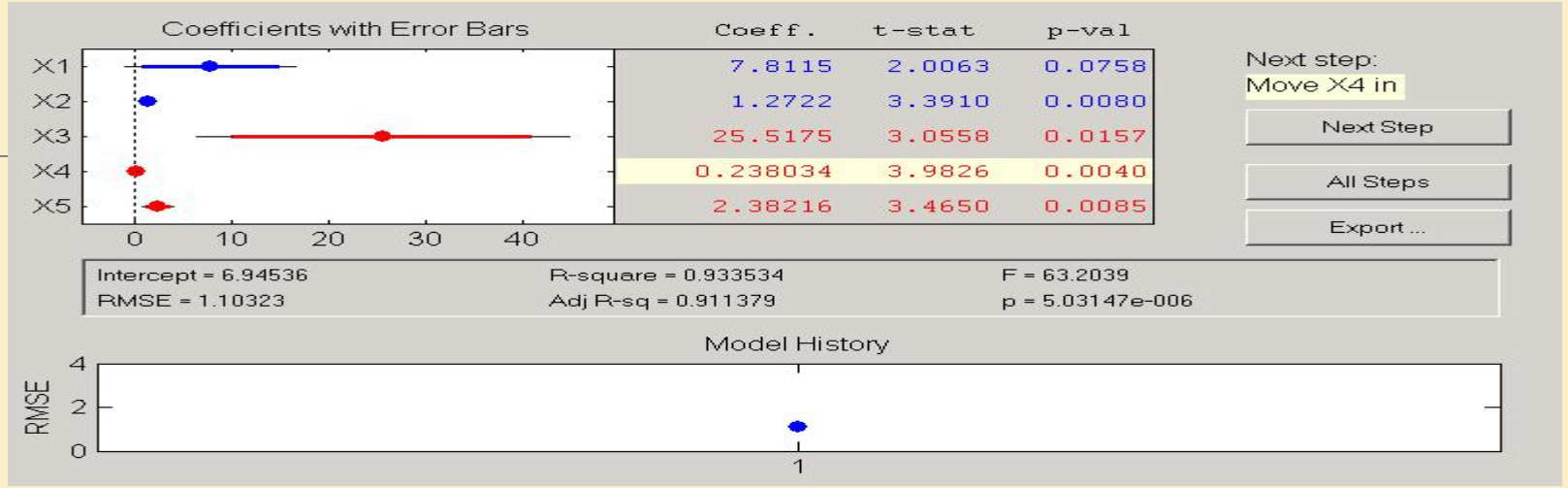


可能存在二次函数关系

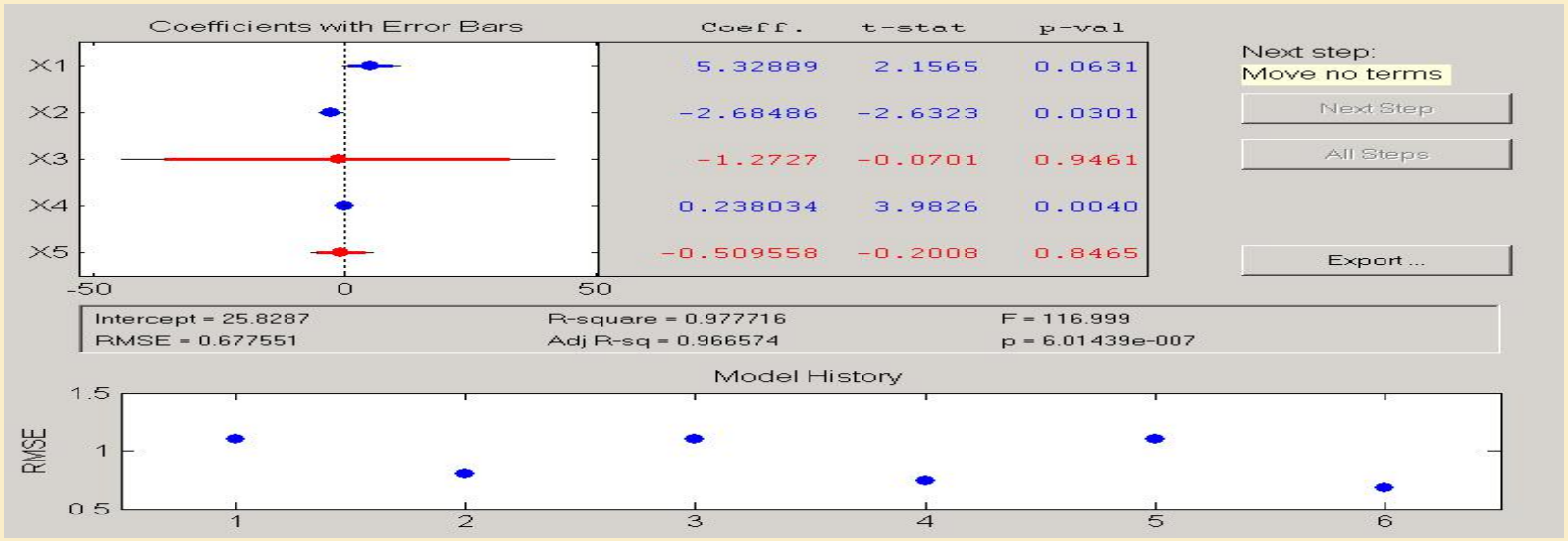
例 儿童的体重与身高和年龄

ertong.m

初始结果



最终结果



$$\hat{y} = 25.8287 + 5.3289\,x_1 - 2.6849\,x_2 + 0.2380\,x_2^2$$

非线性回归分析

非线性最小二乘拟合

已知模型 $y=f(x,\beta),x=(x_1,\cdots,x_m),\beta=(\beta_1,\cdots,\beta_k)$ f对β非线性

观测数据 $(x_i,y_i),\ x_i=(x_{i1},\cdots x_{im}),\ i=1,\cdots,n,\ n>m$

误差平方和 $Q(\beta)=\sum_{i=1}^{n}\varepsilon_i^2(\beta)=\sum_{i=1}^{n}[y_i-f(x_i,\beta)]^2$

非线性回归

$$\begin{cases} y=f(x,\beta)+\varepsilon,\quad x=(x_1,\cdots,x_m),\quad \beta=(\beta_1,\cdots,\beta_k) \\ \varepsilon\sim N(0,\sigma^2) \end{cases}$$

回归系数β的最小二乘估计 $\hat{\beta}$

非线性回归可以对非线性最小二乘拟合结果作统计分析

MATLAB中的非线性回归

[b,R,J]=nlinfit(x,y,'model',b0)

---

x~自变量数据矩阵（每列一个变量）, y~因变量向量,

Model~模型的函数名, m文件: y =f(b,x),b为待估系数β,

b0~回归系数β的初值.

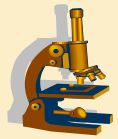
输出：b~β的估计，R~残差，J~估计误差的Jacobi矩阵

bi=nlparci(b,R,J)　　回归系数β的置信区间

nlintool(x,y,'model',b)　一个交互式画面

（内容和用法与多项式回归的Polytool类似）

实例4　酶促反应(续)



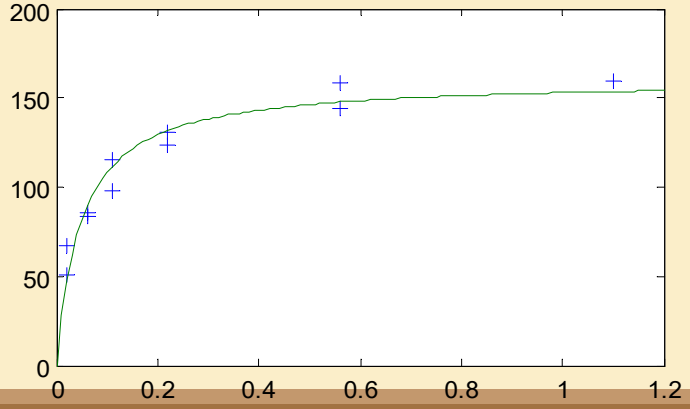
模型　$y=\dfrac{\beta_1 x}{\beta_2+x}$　　y ~ 酶促反应的速度，x ~底物浓度

对未经嘌呤霉素处理的数据　　huaxue1.m

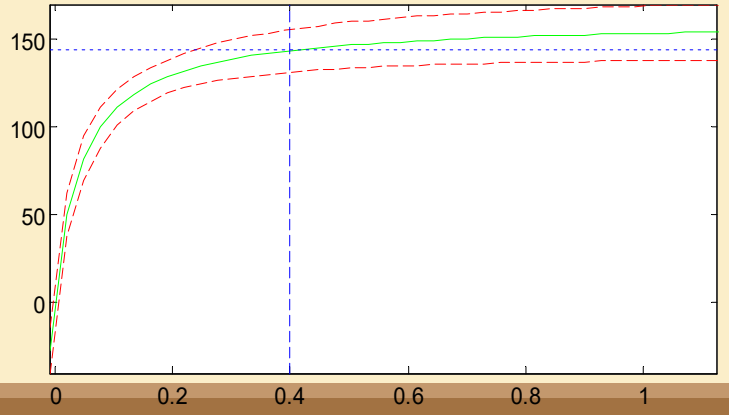
nlinfit ⇨　$\hat{\beta}_1=160.2781,\ \hat{\beta}_2=0.0477,$　与用**lsqnonlin**的结果相同

nlparci ⇨　$\hat{\beta}_1\in[145.6191,174.9372],\ \hat{\beta}_2\in[0.0301,0.0653]$

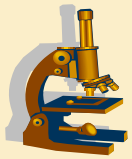
数据拟合结果



nlintool的交互式画面



酶促反应的混合反应模型



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~底物浓度， x_2~ 0(未经处理),1(经过处理)变量,

β_1~未经处理的最终反应速度,

β_2~未经处理的反应的半速度点,

γ_1~经处理后最终反应速度的增长值,

γ_2~经处理后反应的半速度点的增长值 .