第三节　一元线性回归

一、回归分析的基本思想

二、一元线性回归的数学模型

三、可化为一元线性回归的问题

四、小结

一、回归分析的基本思想

变量之间的关系 → 确定性关系

变量之间的关系 → 相 关 关 系

$S=\pi r^2$　　确定性关系

身高和体重　　相关关系

相关关系的特征是:变量之间的关系很难用一种精确的方法表示出来.

确定性关系和相关关系的联系

由于存在测量误差等原因,确定性关系在实际问题中往往通过相关关系表示出来;另一方面,当对事物内部规律了解得更加深刻时,相关关系也有可能转化为确定性关系.

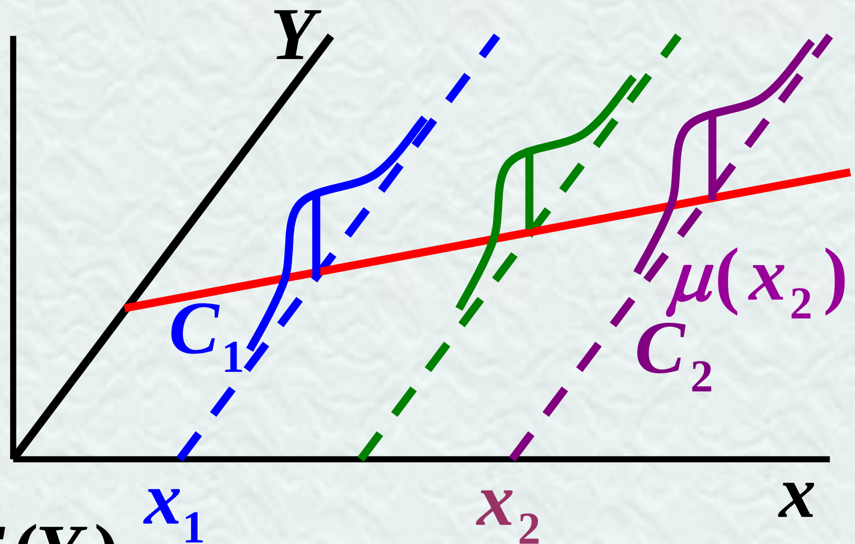
回归分析——处理变量之间的**相关关系**的一种数学方法,它是最常用的数理统计方法.

- 回归分析
  - 线　性回归分析
    - 一元线性回归分析
    - 多元线性回归分析
  - 非线性回归分析

二、一元线性回归的数学模型

问题的分析

设随机变量Y(因变量)和普通变量x(自变量)之间存在着相关关系.



$F(y|x)$表示当x取确定的值x时,所对应的Y的分布函数.

考察Y的数学期望$E(Y)$.

$E(Y)=\mu_{Y|x}=\mu(x)$ Y关于x的回归函数

$$E(Y)=\mu_{Y|x}=\mu(x)$$

因为对随机变量 η, 当 $c=E(\eta)$ 时, $E[(\eta-c)^2]$ 达到最小.

所以在一切 x 的函数中以回归函数 $\mu(x)$ 作为 Y 的近似, 均方误差 $E[(Y-\mu(x))^2]$ 为最小.

实际问题中的 $\mu(x)$ 一般未知.

回归分析的任务——根据试验数据估计回归函数;讨论回归函数中参数的点估计、区间估计;对回归函数中的参数或者回归函数本身进行假设检验;利用回归函数进行预测与控制等等.

问题的一般提法

对 x 的一组不完全相同的值 $x_1, x_2, \cdots, x_n$, 设 $Y_1, Y_2, \cdots, Y_n$ 分别是在 $x_1, x_2, \cdots, x_n$ 处对 Y 的独立观察结果.

称 $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$ 是一个样本.
对应的样本值记为

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n).$$

利用样本来估计 Y 关于 x 的回归函数 $\mu(x)$.

求解步骤

1.推测回归函数的形式

方法一　根据专业知识或者经验公式确定;

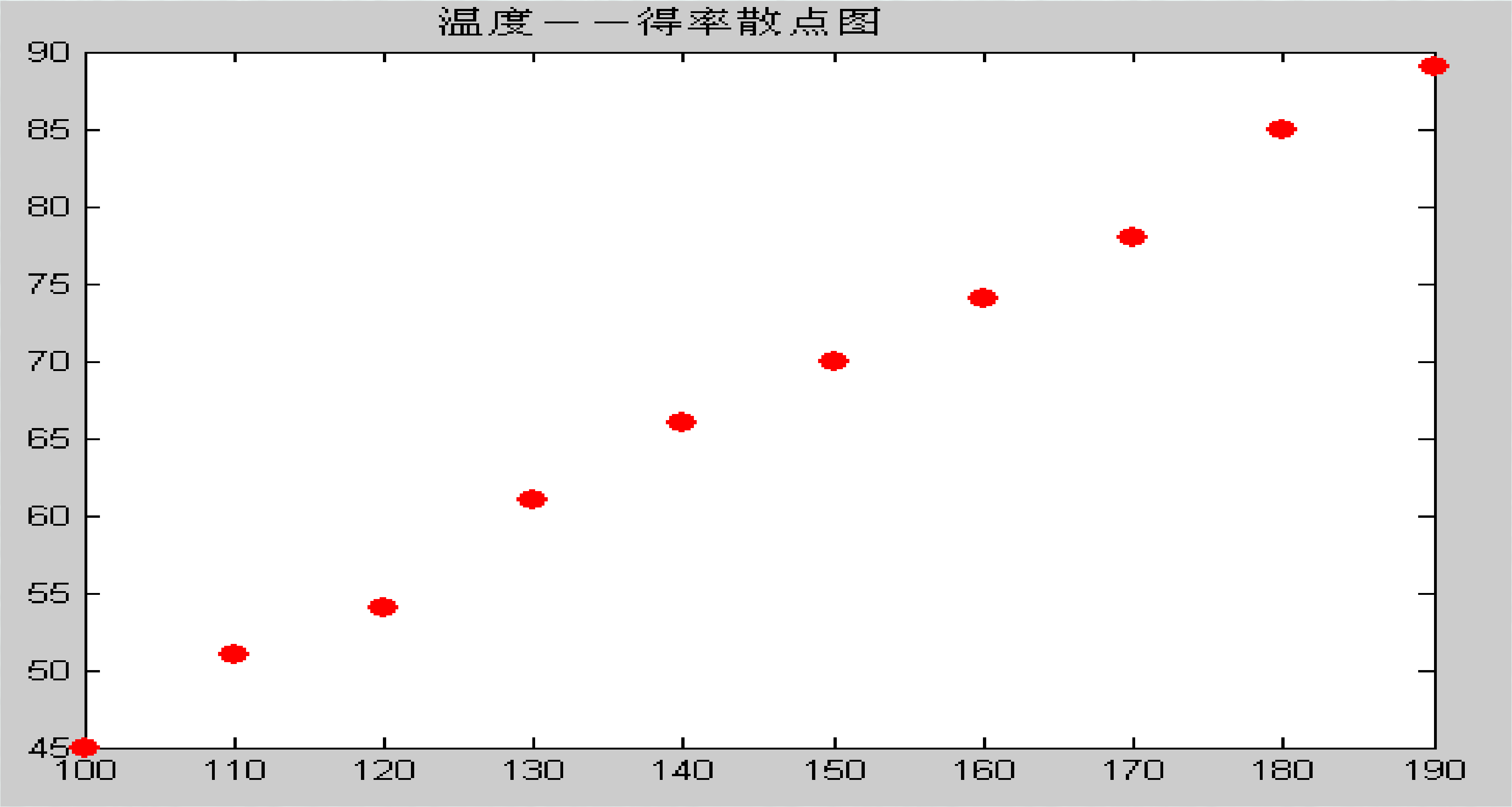
方法二　作散点图观察.

例1　为研究某一化学反应过程中,温度$x(^{o}C)$对产品得率Y (%)的影响, 测得数据如下 .

温度$x(^{o}C)$	100	110	120	130	140	150	160	170	180	190
得率$Y(\%)$	45	51	54	61	66	70	74	78	85	89

用*MATLAB*画出散点图

x=100:10:190;y=[45,51,54,61,66,70,74,78,85,89];
plot(x,y,'.r')

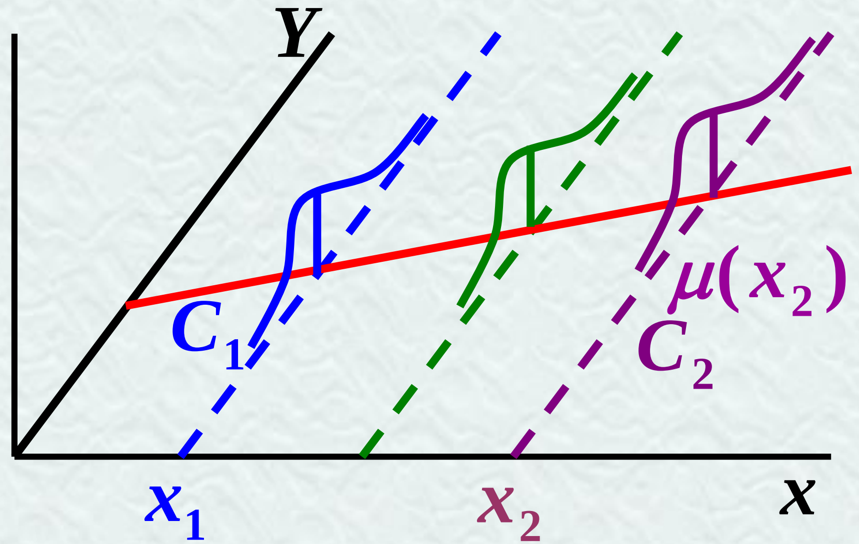


观察散点图，$\mu(x)$具有线性函数$a+bx$的形式.

2.建立回归模型

$\mu(x) = a + bx$ 一元线性回归问题

假设对于x的每一个值有$Y \sim N(a+bx, \sigma^2)$，a，b，σ^2都是不依赖于x的未知参数. $E(Y) = \mu_{Y|x} = \mu(x)$



2.建立回归模型

$\mu(x) = a + bx$　一元线性回归问题

假设对于x的每一个值有$Y \sim N(a+bx, \sigma^2)$, a, b, σ^2都是不依赖于x的未知参数.

记$\varepsilon = Y - (a+bx)$, 那么

$$\boxed{\begin{array}{l} Y = \underbrace{a+bx}_{x\text{的线性函数}} + \underbrace{\varepsilon}_{\text{随机误差}}, \quad \varepsilon \sim N(0, \sigma^2). \\ a, b, \sigma^2 \text{是不依赖于} x \text{的未知参数.} \end{array}}$$

一元线性回归模型

b为回归系数。

3.未知参数a,b的估计

$$Y=a+bx+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2).$$

对于样本 $(x_1,Y_1),(x_2,Y_2),\cdots,(x_n,Y_n)$

$$Y_i=a+bx_i+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma^2),$$ 各 ε_i 相互独立.

于是 $Y_i\sim N(a+bx_i,\sigma^2),\ i=1,2,\cdots,n.$

根据$Y_1,Y_2,\cdots,Y_n$的独立性可得到联合密度函数为

$$L=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y_i-a-bx_i)^2\right]$$

$$=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-a-bx_i)^2\right].$$

用最大似然估计估计未知参数 a, b.

对于任意一组观察值 $y_1, y_2, \cdots, y_n$, 样本的似然函数为

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right]$$

L取最大值等价于

$$Q(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

取最小值.

$$\left.\begin{aligned}\frac{\partial Q}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0\\ \frac{\partial Q}{\partial b} &= -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0\end{aligned}\right\}$$

$$\left.\begin{aligned} na+(\sum_{i=1}^{n}x_i)b &= \sum_{i=1}^{n}y_i \\ (\sum_{i=1}^{n}x_i)a+(\sum_{i=1}^{n}x_i^2)b &= \sum_{i=1}^{n}x_iy_i \end{aligned}\right\}\text{正规方程组}$$

$$\begin{vmatrix} n & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 \end{vmatrix} \neq 0,$$

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

$$\hat{a}=\bar{y}-\hat{b}\bar{x},$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\ \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i.$

$$\mu(x) = a + bx$$

$\hat{\mu}(x) = \hat{a} + \hat{b}x$ Y 关于 x 的经验回归函数

$\hat{y} = \hat{a} + \hat{b}x$ Y 关于 x 的经验回归方程

（简称回归方程）

由于$\hat{a} = \bar{y} - \hat{b}\bar{x}$,

$$\hat{y} = \bar{y} + \hat{b}(x - \bar{x}),$$

回归直线通过散点图的几何中心$(\bar{x}, \bar{y})$.

记 $$S_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2,$$

$$S_{yy}=\sum_{i=1}^{n}(y_i-\overline{y})^2,$$

$$S_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}),$$

$$\hat{b}=\frac{S_{xy}}{S_{xx}}, \qquad \hat{a}=\frac{1}{n}\sum_{i=1}^{n}y_i-\Big(\frac{1}{n}\sum_{i=1}^{n}x_i\Big)\hat{b}.$$

例2 例1中的随机变量 Y 符合一元线性回归模型所述的条件, 求 Y 关于 x 的线性回归方程 . $\hat{y} = \hat{a} + \hat{b}x$

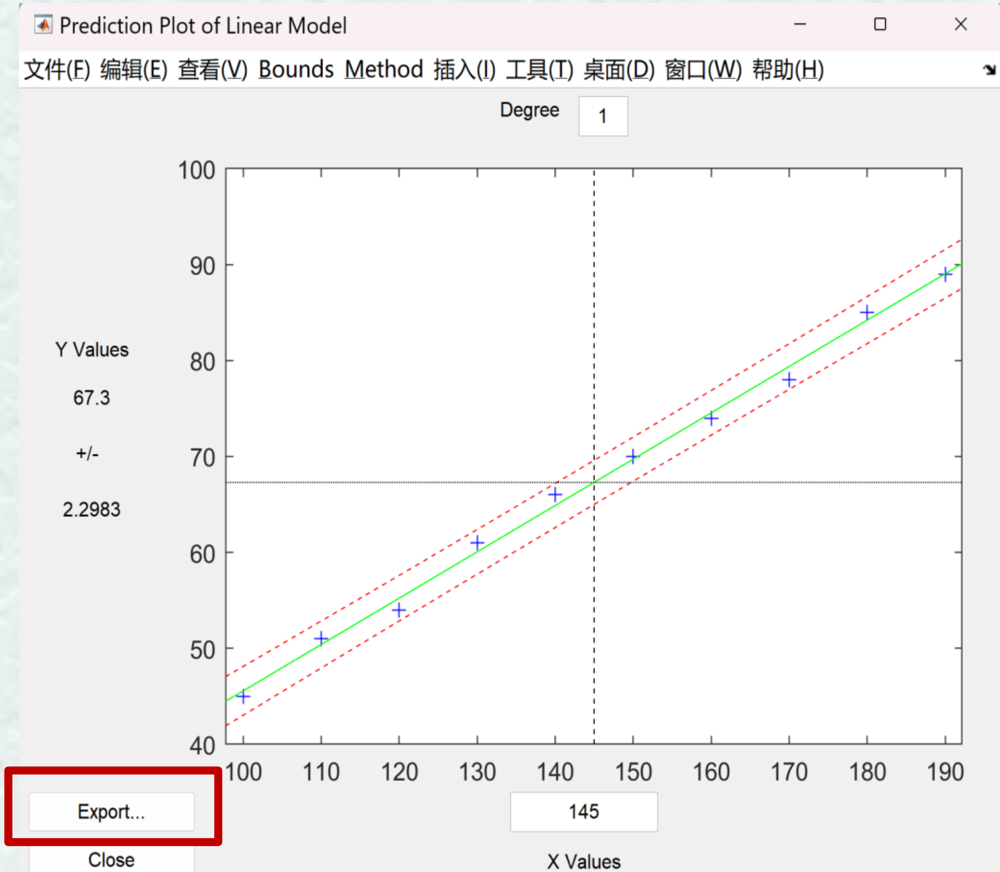
温度x(°C)	100	110	120	130	140	150	160	170	180	190
得率Y(%)	45	51	54	61	66	70	74	78	85	89

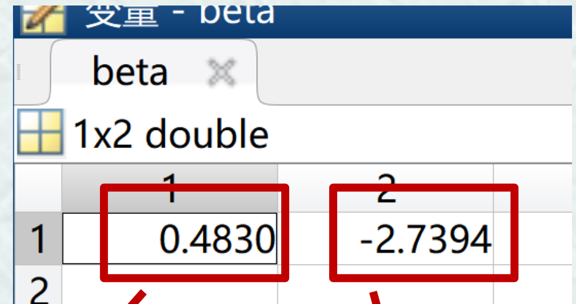
在*MATLAB*中求解

源程序

```
x=100:10:190;
y=[45,51,54,61,66,70,74,78,85,89];
polytool(x,y,1,0.05)
```

程序运行结果 回归图形 参数传送 置信区间 帮 助


Prediction Plot of Linear Model
文件(F) 编辑(E) 查看(V) Bounds Method 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
Degree
1
Y Values
67.3
+/-
2.2983
100
90
80
70
60
50
40
100 110 120 130 140 150 160 170 180 190
Export...
145
Close
X Values


变量 - beta
beta
1x2 double
1
2
1
0.4830
-2.7394
2

$\hat{b}$

$\hat{a}$

4.未知参数σ^2的估计 $\boxed{\mu(x)=a+bx}$

$$Y=a+bx+\varepsilon,\quad \varepsilon \sim N(0,\sigma^2).$$

$$E\{[Y-(a+bx)]^2\}=E(\varepsilon^2)=D(\varepsilon)+[E(\varepsilon)]^2=\sigma^2.$$

σ^2 越小,用回归函数 $\mu(x)=a+bx$ 作为Y 的近似导致的均方误差就越小.

$$\hat{y}_i=\hat{y}\big|_{x=x_i}=\hat{a}+\hat{b}x_i,$$

$y_i-\hat{y}_i$ x_i处的残差

$$Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2$$

残差平方和

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - \bar{y} - \hat{b}(x_i - \bar{x})]^2$$

$$= \cdots = \boxed{S_{yy} - \hat{b}S_{xy}}.$$

记 $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, $S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$.

b, a 的估计量为 $\hat{b} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\hat{a} = \bar{Y} - \hat{b}\bar{x}$

残差平方和 Q_e 的相应的统计量为

$$Q_e = S_{YY} - \hat{b}S_{xY}.$$

残差平方和 Q_e 的相应的统计量为

$$Q_e = S_{YY} - \hat{b} S_{xY}.$$

可以证明 $\dfrac{Q_e}{\sigma^2} \sim \chi^2(n-2),$

从而 $E\left(\dfrac{Q_e}{\sigma^2}\right) = n-2, E\left(\dfrac{Q_e}{n-2}\right) = \sigma^2.$

σ^2 的无偏估计量为

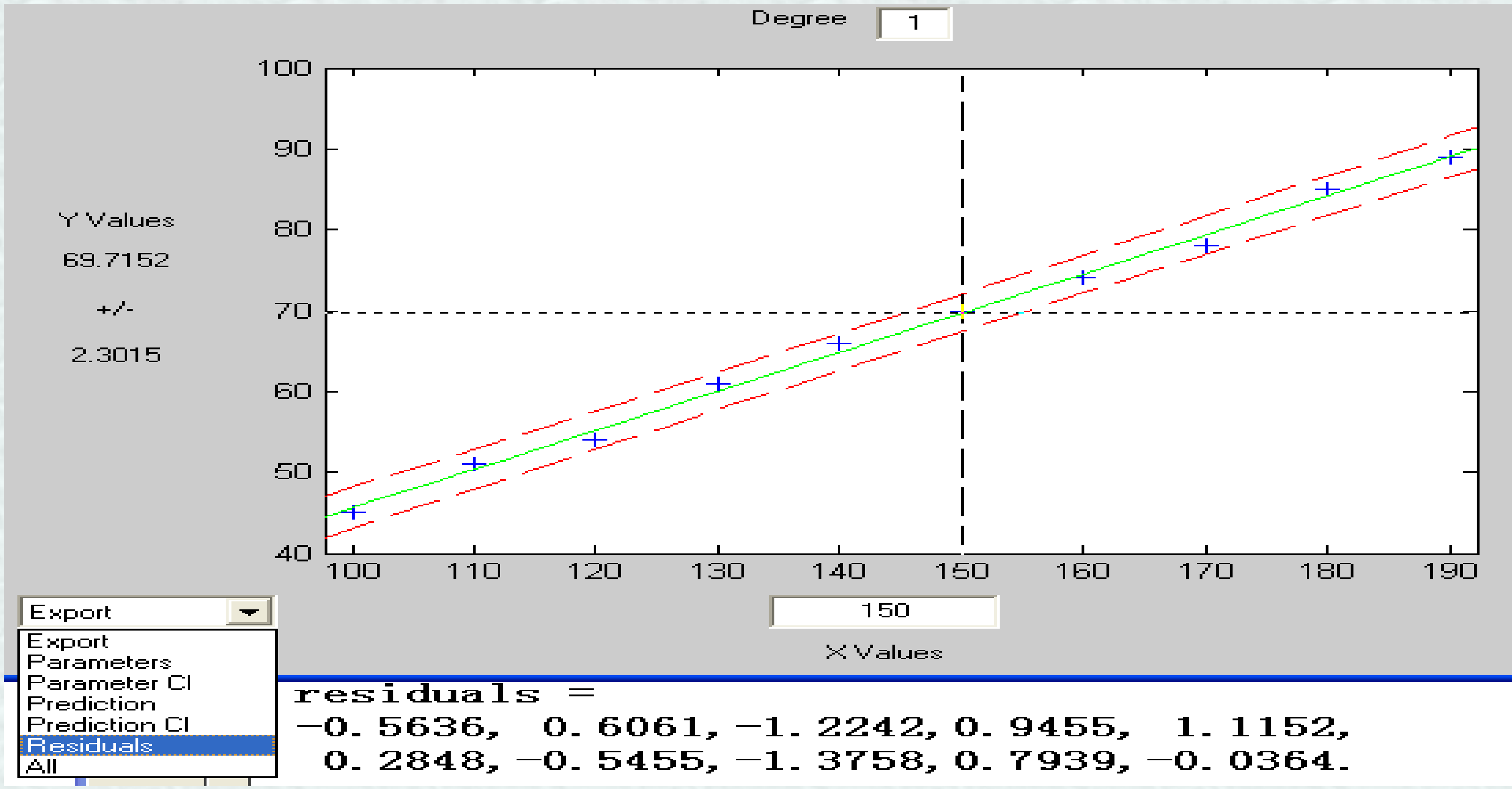
$$\hat{\sigma}^2 = \frac{Q_e}{n-2} = \frac{1}{n-2}\left[S_{YY} - \hat{b} S_{xY}\right].$$

例3 求例2中方差的无偏估计.

例2 例1中的随机变量 Y 符合一元线性回归模型所述的条件, 求 Y 关于 x 的线性回归方程 .

温度x(°C)	100	110	120	130	140	150	160	170	180	190
得率Y(%)	45	51	54	61	66	70	74	78	85	89

例3 求例2中方差的无偏估计.



$$Q_e = \sum_{i=1}^{10}(residuals)_i^2 = 7.2236, \hat{\sigma}^2 = \frac{7.2236}{8} = 0.9030.$$

5.线性假设的显著性检验

$$Y = a + bx + \varepsilon, \quad \varepsilon \sim N(0,\sigma^2).$$

检验假设： $H_0 : b = 0, \quad H_1 : b \neq 0.$

$\boxed{\hat{b} \sim N(b, \sigma^2 / S_{xx}),}$ $\quad \dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} = \dfrac{Q_e}{\sigma^2} \sim \chi^2(n-2).$

并且$\hat{b}, Q_e$相互独立,因此

$$\frac{\hat{b} - b}{\hat{\sigma}} \sqrt{S_{xx}} \sim t(n-2).$$

当 H_0 为真时 $b = 0$, 此时 $t = \dfrac{\hat{b}}{\hat{\sigma}} \sqrt{S_{xx}} \sim t(n-2)$,

并且$E(\hat{b}) = b = 0$,得H_0的拒绝域为

$$|t| = \frac{|\hat{b}|}{\hat{\sigma}} \sqrt{S_{xx}} \geq t_{\alpha/2}(n-2).$$

拒绝 $H_0: b=0$, 认为回归效果显著.

接受 $H_0: b=0$, 认为回归效果不显著.

回归效果不显著的原因分析:

(1) 影响 Y 取值的, 除 x 及随机误差外还有其他不可忽略的因素;

(2) $E(Y)$ 与 x 的关系不是线性的;

(3) Y 与 x 不存在关系.

例4 检验例 2 中的回归效果是否显著,取显著性水平为 0.05 .

解 已知

$$\hat{b} = 0.4830,\ S_{xx} = 8250,\ \hat{\sigma}^2 = 0.9030,$$

查表得 $t_{0.05/2}(n-2) = t_{0.025}(8) = 2.3060.$

$$|t| = \frac{0.4830}{\sqrt{0.9030}} \times \sqrt{8250} = 46.25,$$

$$|t| > t_{0.025}(8).$$

拒绝 $H_0 : b = 0$, 认为回归效果显著 .

6.系数b的置信区间

当回归效果显著时，对系数 b 作区间估计.

$$\hat{b} \sim N(b, \sigma^2 / S_{xx}), \qquad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{Q_e}{\sigma^2} \sim \chi^2(n-2).$$

$$\frac{\hat{b} - b}{\hat{\sigma}} \sqrt{S_{xx}} \sim t(n-2).$$

系数 b 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\hat{b} \pm t_{\alpha/2}(n-2) \times \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \right).$$

6.系数b的置信区间

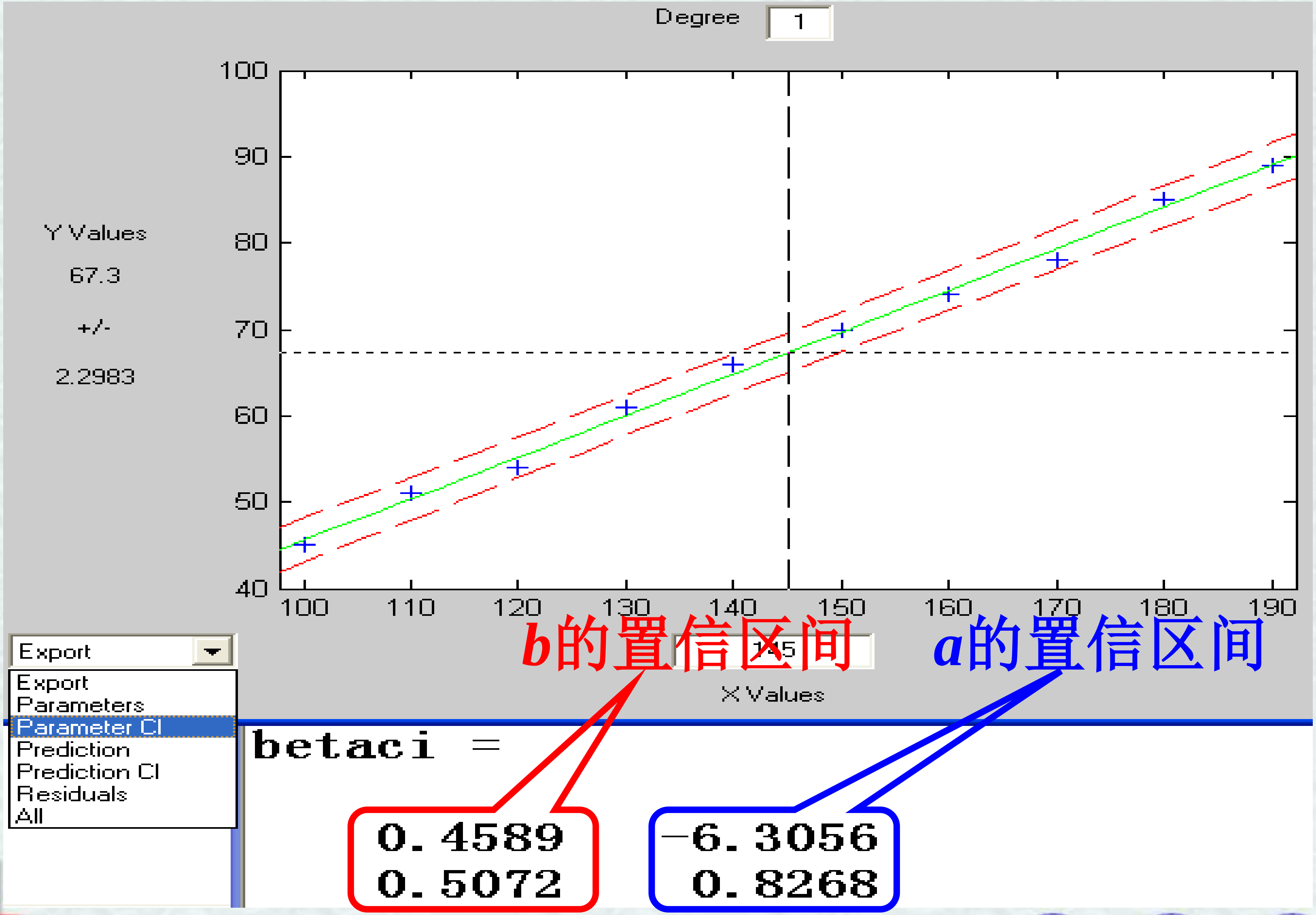
当回归效果显著时,对系数 b 作区间估计.

系数 b 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\hat{b} \pm t_{\alpha/2}(n-2) \times \frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right).$$

例如,求例1中 b 的置信水平为 0.95 的置信区间.

$$\left(0.4830 \pm 2.3060 \times \sqrt{\frac{0.9030}{8250}}\right) = (0.45894, 0.50712).$$


Degree
1
Y Values
67.3
+/-
2.2983
100
90
80
70
60
50
40
100
110
120
130
140
150
160
170
180
190
X Values
145
Export
Export
Parameters
Parameter CI
Prediction
Prediction CI
Residuals
All
b的置信区间
a的置信区间
betaci =
0.4589
0.5072
-6.3056
0.8268

7.回归函数函数值的点估计和置信区间

经验回归函数（回归函数的估计）

$$\hat{y} = \hat{\mu}(x) = \hat{a} + \hat{b}x,$$

$x = x_0$，估计值 $\hat{y}_0 = \hat{\mu}(x_0) = \hat{a} + \hat{b}x_0$，

估计量：$\hat{Y}_0 = \hat{a} + \hat{b}x_0$.

因为 $\boxed{E(\hat{Y}_0) = a + bx_0\text{，所以估计量是无偏的}}$.

（见教材附录）

7.回归函数函数值的点估计和置信区间

$$\frac{\hat{Y}_0-(a+bx_0)}{\sigma\sqrt{\dfrac{1}{n}+\dfrac{(x_0-\overline{x})^2}{S_{xx}}}}\sim N(0,1),$$

$Q_e, \hat{Y}_0$相互独立.

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2}=\frac{Q_e}{\sigma^2}\sim\chi^2(n-2),$$

$$\frac{\hat{Y}_0-(a+bx_0)}{\hat{\sigma}\sqrt{\dfrac{1}{n}+\dfrac{(x_0-\overline{x})^2}{S_{xx}}}}\sim t(n-2),$$

$$\frac{\hat{Y}_0-(a+bx_0)}{\hat{\sigma}\sqrt{\dfrac{1}{n}+\dfrac{(x_0-\overline{x})^2}{S_{xx}}}}\sim t(n-2),$$

回归函数值

$\mu(x_0)=a+bx_0$ 的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\hat{Y}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}\right)$$

或

$$\left(\hat{a}+\hat{b}x_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}\right).$$

8. Y 的观察值的点预测和预测区间

设 Y_0 是在 $x = x_0$ 处对 Y 的观察结果.

$$Y_0 = a + bx_0 + \varepsilon_0, \quad \varepsilon_0 \sim N(0, \sigma^2).$$

利用 $x = x_0$ 处经验回归函数的函数值作为 Y_0 的点预测

$$\hat{Y}_0 = \hat{\mu}(x_0) = \hat{a} + \hat{b}x_0 \quad Y_0\text{的点预测}$$

8. Y 的观察值的点预测和预测区间

给定置信水平为 $1-\alpha$,

Y_0 的置信水平为 $1-\alpha$ 的预测区间

$$\left(\hat{Y}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}\right)$$

或

$$\left(\hat{a}+\hat{b}x_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}\right)$$

例5 (续例2)

(1) 求回归函数 $\mu(x)$ 在 $x = 125$ 处的值 $\mu(125)$ 的置信水平为 0.95 的置信区间, 求在 $x = 125$ 处 Y 的新观察值 Y_0 的置信水平为 0.95 的预测区间;

回归函数值

$\mu(x_0) = a + bx_0$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\hat{Y}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \right)$$

解　(1)已知

$$\hat{b} = 0.4830,\ \hat{a} = -2.7394,\ S_{xx} = 8250,$$

$$\hat{\sigma}^2 = 0.9030,\ \overline{x} = 145.$$

查表得 $t_{0.05/2}(n-2)=t_{0.025}(8)=2.3060.$

计算 $\hat{Y}_0=\hat{Y}\big|_{x=125}=57.64,$

$$t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}=0.84,$$ 回归函数值$\mu(x_0)$

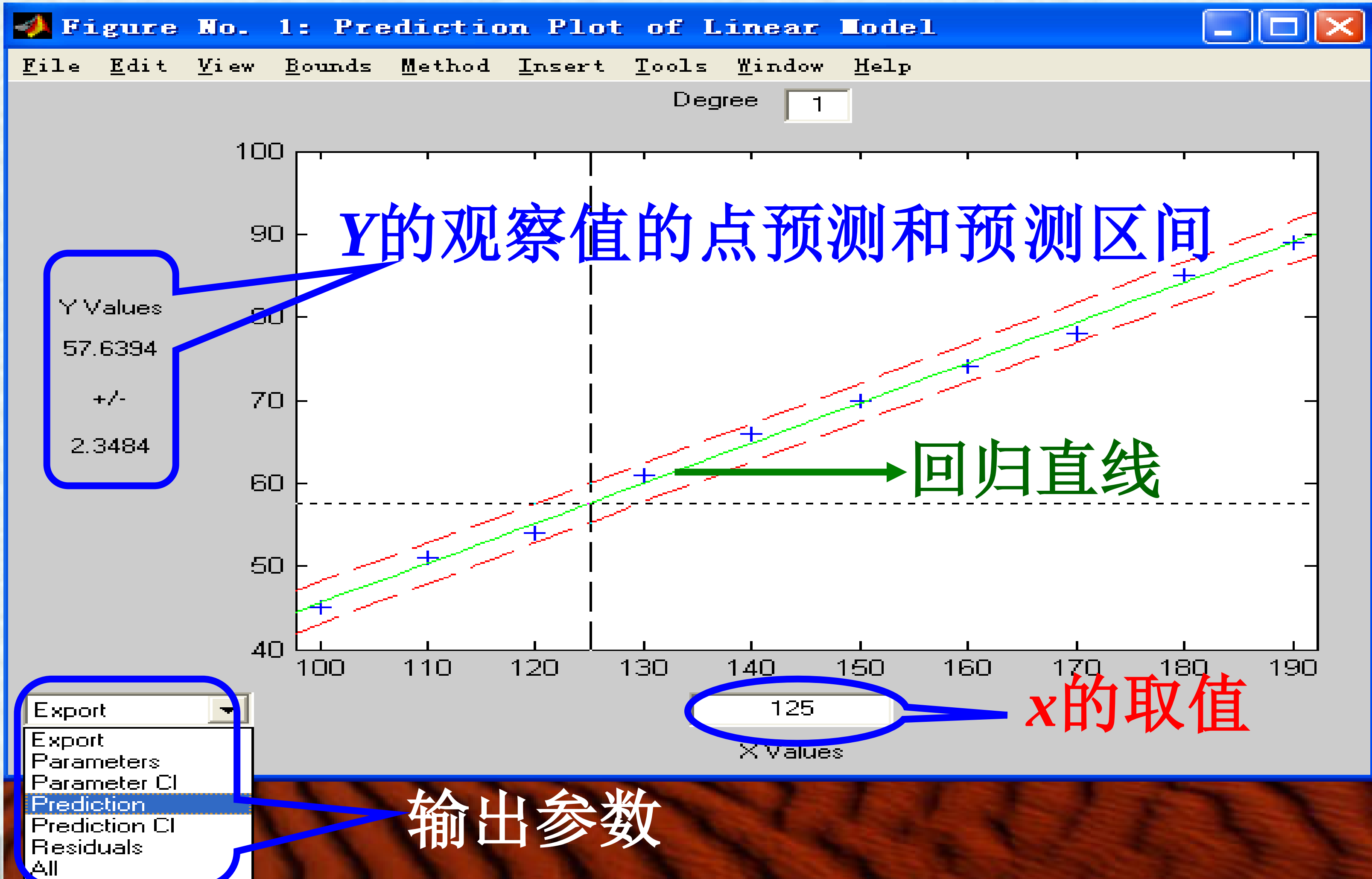
$$t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}=2.34.$$ 观察值Y_0

回归函数 $\mu(x)$ 在 $x=125$ 处的值 $\mu(125)$ 的置信水平为 0.95 的置信区间为 (57.64 ± 0.84).

在 $x_0=125$ 处 Y 的新观察值 Y_0 的置信水平为 0.95 的预测区间为 (57.64 ± 2.34).

(2) 求在 $x = x_0$ 处 Y 的新观察值 Y_0 的置信水平为 0.95 的预测区间.

(2)在*MATLAB*中求解



三、可化为一元线性回归的问题

方法——通过适当的变量变换,化成一元线性回归问题进行分析处理.

1. $Y = \alpha e^{\beta x} \bullet \varepsilon, \quad \ln\varepsilon \sim N(0,\sigma^2).$

两边取对数

$$\ln Y = \ln\alpha + \beta x + \ln\varepsilon.$$

$$\Downarrow$$

$$Y' = a + bx' + \varepsilon', \quad \varepsilon' \sim N(0,\sigma^2).$$

$$2.\ Y=\alpha x^{\beta}\bullet\varepsilon,\quad \ln\varepsilon\sim N(0,\sigma^2).$$

两边取对数

$$\ln Y=\ln\alpha+\beta\ln x+\ln\varepsilon.$$

$$\Downarrow$$

$$Y'=a+bx'+\varepsilon',\quad \varepsilon'\sim N(0,\sigma^2).$$

$$3.\ Y=\alpha+\beta h(x)+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2).$$

$$\Downarrow$$

$$Y=a+bx'+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2).$$

例6 表 9.18 是 1957 年美国旧轿车价格的调查资料，今以 x 表示轿车的使用年数，Y 表示相应的平均价格(以美元计)，求 Y 关于 x 的回归方程 .

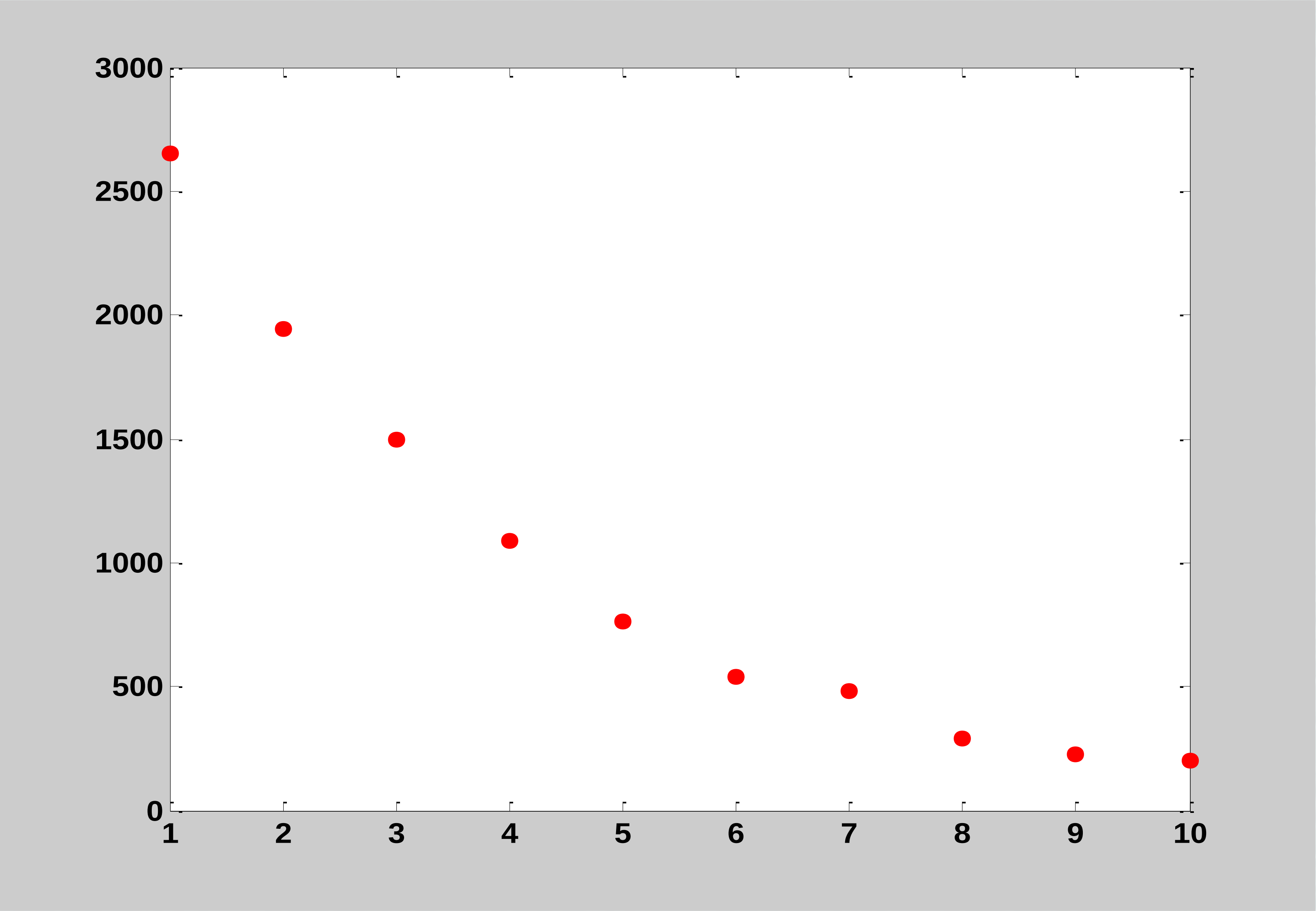
表 9.18

年数x	1	2	3	4	5	6	7	8	9	10
价格Y	2651	1943	1494	1087	765	538	484	290	226	204

在*MATLAB*中求解

首先作散点图

```
x=1:1:10;
y=[2651,1943,1494,1087,765,538,484,290,226,204];
plot(x,y,'.r')
```


3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10

选择模型

$$Y=\alpha e^{\beta x}\bullet\varepsilon,\quad \ln\varepsilon\sim N(0,\sigma^2).$$

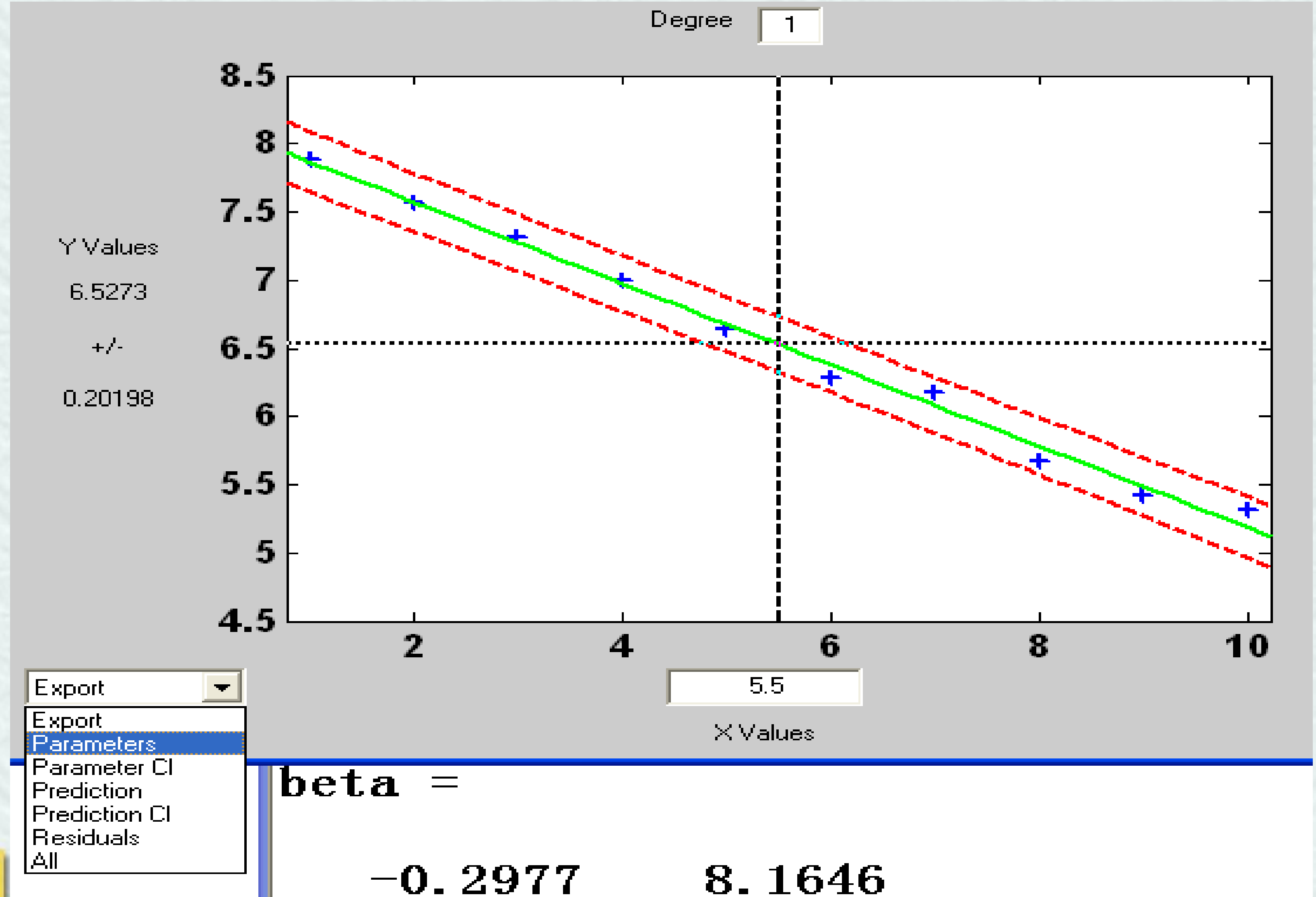
变量变换

令 $\ln Y=Y',\ln\alpha=a,\beta=b,x=x',\ln\varepsilon=\varepsilon'.$

$$Y'=a+bx'+\varepsilon',\quad \varepsilon'\sim N(0,\sigma^2).$$

数据变换 xx=x;yy=ln(y);

求回归方程 polytool(xx,yy,1)


Degree
1
8.5
8
7.5
7
6.5
6
5.5
5
4.5
Y Values
6.5273
+/-
0.20198
2
4
6
8
10
5.5
X Values
Export
Export
Parameters
Parameter CI
Prediction
Prediction CI
Residuals
All
beta =
-0.2977 8.1646

$$\hat{b} = -0.2977, \hat{a} = 8.1646.$$

$$\hat{y}' = -0.2977x + 8.1646.$$

线性假设的显著性检验

$$|t| = \frac{\hat{b}}{\hat{\sigma}}\sqrt{S_{xx}} = 32.3693 > t_{0.05/2}(8) = 2.3060.$$

线性回归效果高度显著.

代回原变量,得曲线回归方程

$$\hat{y} = \exp(\hat{y}') = \exp(-0.2977x + 8.1646)$$

$$= 3514.3e^{-0.2977x}.$$

四、小结

1.回归分析的任务

研究变量之间的相关关系

2.一元线性回归的步骤

(1) 推测回归函数; (2) 建立回归模型;
(3) 估计未知参数; (4) 进行假设检验;
(5) 预测与控制.

3.可化为一元线性回归的问题

关键:选择适当的变量代换.