第五讲　回归分析

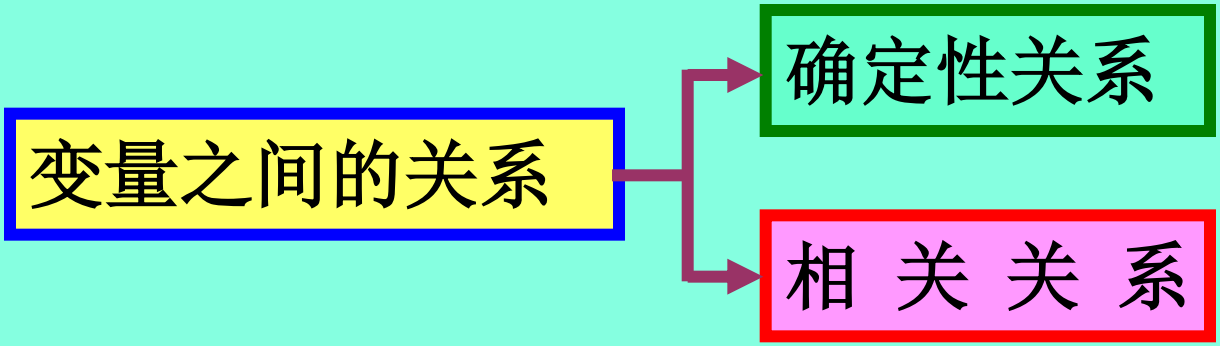
一、线性回归分析的基本的思想

二、一元线性回归

三、可化为一元线性回归的问题

四、多元线性回归

一、回归分析的基本思想



$S = \pi r^2$　　确定性关系

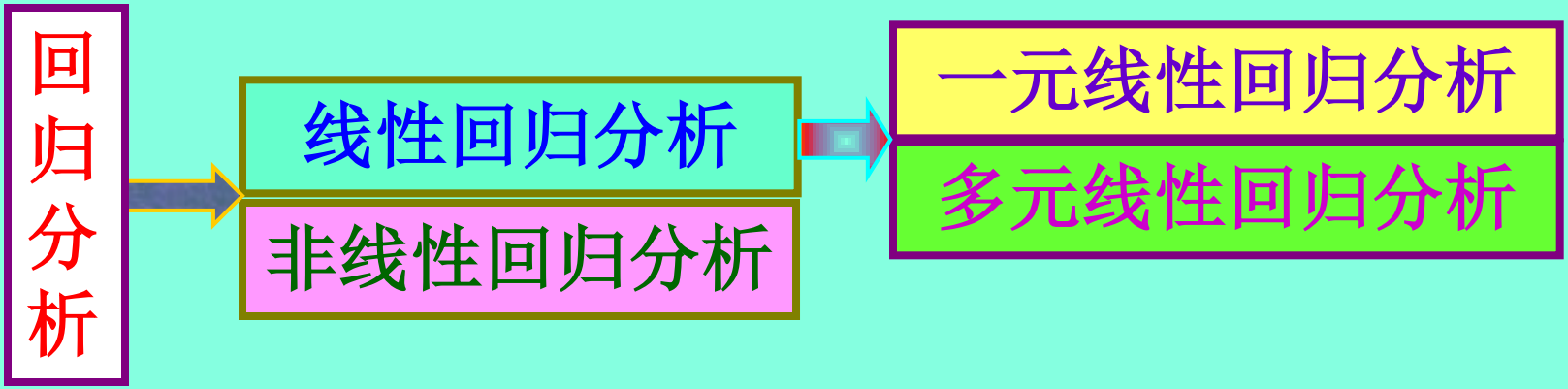
身高和体重　　相关关系

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.

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.



回归分析的任务:

(1)根据试验数据估计回归函数;

(2)讨论回归函数中参数的点估计、区间估计;

(3)对回归函数中的参数或者回归函数本身进行假设检验;

(4)利用回归函数进行预测与控制等等.

特别对随机变量Y的观察值做出点预测和区间预测.

二、一元线性回归

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

可将每对观察值(x_i, y_i)在直角坐标系中描出它的相应的点，这种图称为散点图.

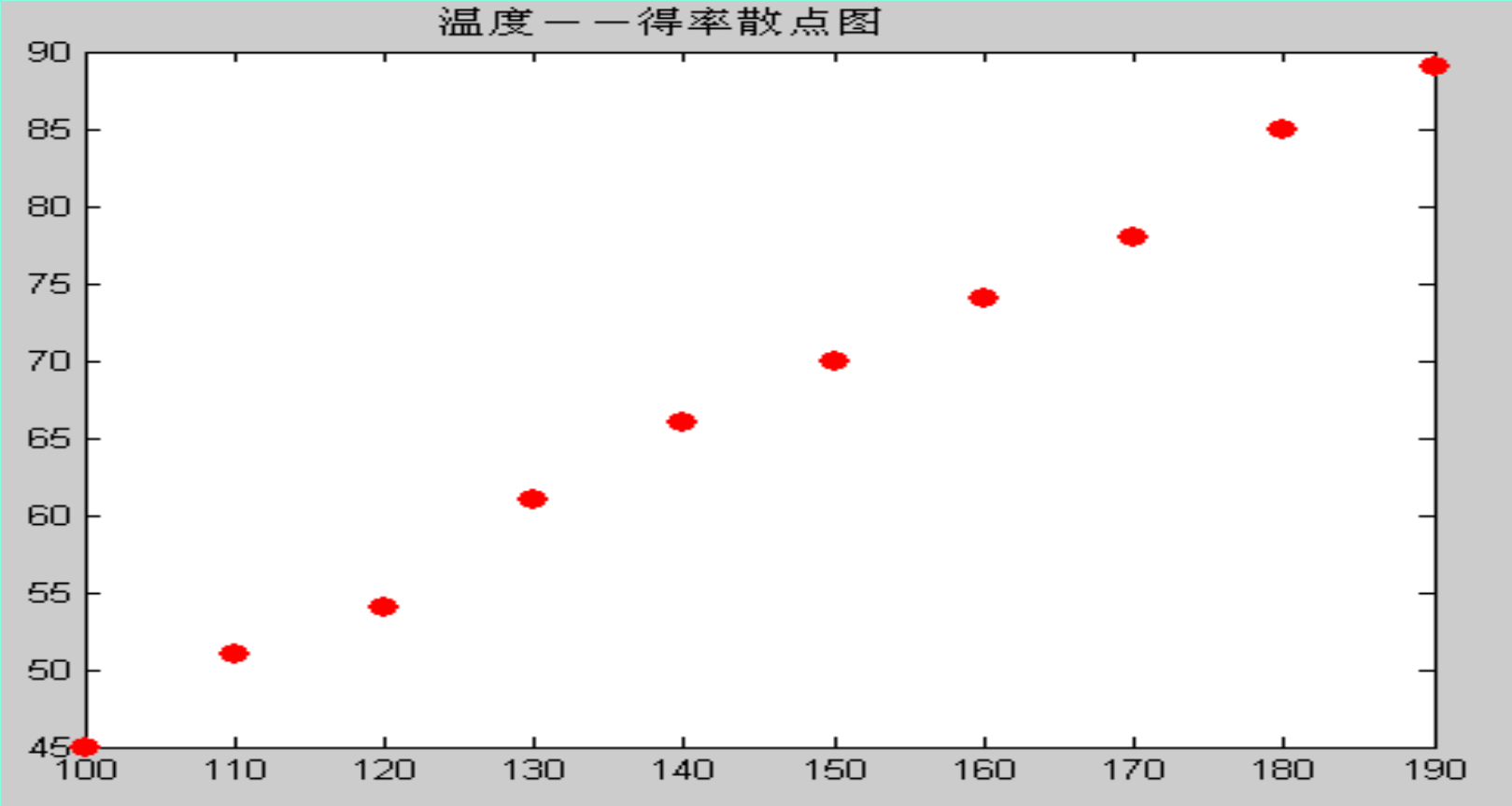
记 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$

例1 为研究某一化学反应过程中, 温度$x(^0C)$对产品得率Y(%)的影响, 测得数据如下 .

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

这里自变量x是普通变量, Y是随机变量.

画出散点图如下,



观察散点图, Y'具有线性函数$a+bx$的形式 .

1. 回归模型

$Y' = a + bx$　　理论线性回归方程

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

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

a, b, σ^2是不依赖于x的未知参数.

一元线性回归模型

x的线性函数　随机误差

假设对于x(在某个区间内)的每一个值

$$Y\sim N(a+bx,\sigma^2)$$

其中a,b,σ^2都是不依赖于x的未知参数. 记

$$\varepsilon=Y-(a+bx)$$

对Y作这样的正态假设，相当于假设

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

其中未知参数a,b及σ^2都不依赖于x .上式称为一元线性回归模型, 其中b称为回归系数 .

注意：

1)Y为随机变量，x为可控变量.即y_i为样本值，x_i可控变量值；

2)$Y' = a+bx \Leftrightarrow y_i' = a+bx_i$为理论线性回归方程；

3)$Y \sim N(a+bx,\sigma^2) \Leftrightarrow Y_i \sim N(a+bx_i,\sigma^2)$,

即$EY_i = a+bx_i, DY_i = \sigma^2$;

4）$E\bar{Y} = a+b\bar{x}, D\bar{Y} = \dfrac{\sigma^2}{n}$.

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

2. 未知参数估计

1）. 未知参数a,b的估计—最小二乘估计法

取x的n个不全相同的值$x_1, x_2, \cdots, x_n$做独立试验，得到样本$(x_1,Y_1),(x_2,Y_2),\cdots,(x_n,Y_n)$.

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

于是$Y_i \sim N(a+bx_i,\sigma^2), i = 1, 2, \cdots, n$. 由$Y_1, Y_2, \cdots, Y_n$的独立性，

即只需函数

$$Q(a,b)=\sum_{i=1}^{n}(y_i-y_i')^2=\sum_{i=1}^{n}\varepsilon_i^{\,2}=\sum_{i=1}^{n}(y_i-a-bx_i)^2$$取最小值.

根据

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

得方程组

$$\begin{cases} na+(\sum_{i=1}^{n}x_i)b=\sum_{i=1}^{n}y_i \\ (\sum_{i=1}^{n}x_i)a+(\sum_{i=1}^{n}x_i^2)b=\sum_{i=1}^{n}x_iy_i \end{cases}$$

由于x_i不全相同，方程组的系数行列式

$$\begin{vmatrix} n & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 \end{vmatrix} = n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2=n\sum_{i=1}^{n}(x_i-\bar{x})^2\neq 0$$

$$\therefore \hat{b} = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} =$$

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

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

$$\hat{a}=\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{\hat{b}}{n}\sum_{i=1}^{n}x_i=\bar{y}-\hat{b}\bar{x}$$

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

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

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

$$S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}(x_iy_i-x_i\bar{y}-y_i\bar{x}+\bar{x}\cdot\bar{y})=$$

$$\sum_{i=1}^{n}x_iy_i-n\bar{x}\cdot\bar{y}=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)=\sum_{i=1}^{n}y_i(x_i-\bar{x})=$$

$$\sum_{i=1}^{n}x_i(y_i-\bar{y}),$$

$$\text{故}\hat{b}=\frac{S_{xy}}{S_{xx}}=\frac{\sum_{i=1}^{n}x_iy_i-\frac{1}{n}(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{\sum_{i=1}^{n}x_i^2-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2},$$

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

取$\hat{a}+\hat{b}x$作为回归函数$Y'=a+bx$的估计，即

$$\hat{Y}=\hat{a}+\hat{b}x$$

称为Y 关于 x 的经验回归函数,记$\hat{a}+\hat{b}x=\hat{y}$，方程

$$\hat{y}=\hat{a}+\hat{b}x \Leftrightarrow \hat{y}_i=\hat{a}+\hat{b}x_i$$

称为Y 关于 x 的经验回归方程,简称回归方程，其图形称为回归直线.

$$\hat{a}=\overline{y}-\hat{b}\overline{x}$$

$$\hat{y}=\overline{y}+\hat{b}(x-\overline{x}) \Leftrightarrow \hat{y}_i=\overline{y}+\hat{b}(x_i-\overline{x}).$$

对于样本值$(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)$，回归直线通过散点图的几何中心$(\overline{x},\overline{y})$.

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 .

	x	y	x^2	y^2	xy
	100	45	10000	2025	4500
	110	51	12100	2601	5610
	120	54	14400	2916	6480
	130	61	16900	3721	7930
	140	66	19600	4356	9240
	150	70	22500	4900	10500
	160	74	25600	5476	11840
	170	78	28900	6084	13260
	180	85	32400	7225	15300
	190	89	36100	7921	16910
$\sum$	1450	673	218500	47225	101570

$$S_{xx}=218500-\frac{1}{10}\times 1450^2=8250$$

$$S_{xy}=101570-\frac{1}{10}\times 1450\times 673=3985$$

$$\hat{b}=S_{xy}/S_{xx}=0.48303$$

$$\hat{a}=\frac{1}{10}\times 673-\frac{1}{10}\times 1450\times 0.48303=-2.73935$$

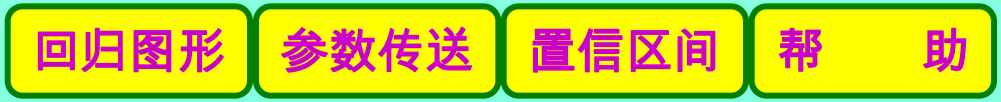
回归直线方程 $\hat{y}=-2.73935+0.48303x$

或 $\hat{y}=67.3+0.48303(x-145)$

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



2）未知参数σ^2的估计

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

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

$$Y'=a+bx$$

去研究随机变量Y与x的关系就愈有效

为了估计σ^2, 引入残差平方和

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

$y_i-\hat{y}_i$为x_i处的残差.

残差平方和

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

是经验回归函数在x_i处的函数值 $\hat{Y} = \hat{a} + \hat{b}x$

与x_i处的观察值y_i的偏差的平方和.

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad \Leftrightarrow \hat{y}_i = \bar{y} + \hat{b}(x_i - \bar{x})$$

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

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

$$= S_{yy} - 2\hat{b}S_{xy} + (\hat{b})^2 S_{xx}$$

由 $\hat{b} = S_{xy} / S_{xx}$

$$Q_e = S_{yy} - 2\hat{b}S_{xy} + \hat{b}S_{xy} = S_{yy} - \hat{b}S_{xy} = S_{yy} - \hat{b}^2 S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

b, a 的估计量为

$$\hat{b} = \frac{S_{xy}}{S_{xx}},$$

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\hat{b}}{n}\sum_{i=1}^{n} x_i = \overline{Y} - \hat{b}\overline{x}$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$.在$S_{yy},S_{xy}$的表达式中

将y_i改为$Y_i(i=1,2,\cdots,n)$,并把它们分别记为S_{YY},S_{xY}

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

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

$$\hat{\sigma}^2=\frac{Q_e}{n-2}$$

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服从分布

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

$$E(\frac{Q_e}{\sigma^2}) = n-2, \qquad E(\frac{Q_e}{n-2}) = \sigma^2.$$ 为什么？

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.

解　$$S_{yy}=\sum_{i=1}^{n}y_i^2-\frac{1}{n}(\sum_{i=1}^{n}y_i)^2$$

$$=47225-\frac{1}{10}\times 673^2$$

$$=1932.1$$

又知$S_{xy}=3985$, $\hat{b}=0.48303$

$$Q_e=S_{yy}-\hat{b}S_{xy}=7.23$$

$$\hat{\sigma}^2=Q_e/(n-2)=7.23/8=0.90.$$

3、$\hat{a}$、$\hat{b}$、$\hat{\sigma}^2$的统计特性

1)$\hat{a}$、$\hat{b}$、$\hat{\sigma}^2$均为a、b、σ^2得无偏估计，

且(1)$\mathrm{cov}(\bar{y},\hat{b})=0$,

(2)$D(\hat{b})=\dfrac{\sigma^2}{S_{xx}}$,

(3)$D(\hat{a})=\sigma^2(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}})$,

(4)$\mathrm{cov}(\hat{a},\hat{b})=-\dfrac{\bar{x}\sigma^2}{S_{xx}}$.　(5)$E(\dfrac{Q_e}{n-2})=\sigma^2$.

2)$\hat{b}\sim N(b,\dfrac{\sigma^2}{S_{xx}})$, $\hat{a}\sim N(a,\sigma^2(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}))$.

$$1)E\hat{b} = E(\frac{S_{xy}}{S_{xx}}) = \frac{E(\sum_{i=1}^{n}(x_i - \bar{x})y_i)}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Ey_i}{S_{xx}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(a + bx_i)}{S_{xx}} = \frac{a\sum_{i=1}^{n}(x_i - \bar{x}) + b\sum_{i=1}^{n}(x_i - \bar{x})x_i}{S_{xx}}$$

$$= \frac{a(\sum_{i=1}^{n}x_i - n\bar{x}) + b\sum_{i=1}^{n}(x_i - \bar{x})x_i}{S_{xx}} = \frac{bS_{xx}}{S_{xx}} = b.$$

$$E\hat{a} = E(\bar{y} - \hat{b}\bar{x})$$

$$= E\bar{y} - \bar{x}E\hat{b} = a + b\bar{x} - b\bar{x} = a.$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}}\operatorname{cov}(y_i,y_i) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}} =$$

$$\frac{\sigma^2(\sum_{i=1}^{n}x_i - n\bar{x})}{nS_{xx}} = 0.$$

$$(2) D(\hat{b}) = D\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{S_{xx}}\right)$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Dy_i}{S_{xx}^2} = \frac{\sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{S_{xx}^2}$$

$$= \frac{\sigma^2 S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.$$

$$(3)D(\hat{a}) = D(\bar{y} - \hat{b}\bar{x}) = D\bar{y} + \bar{x}^2 D\hat{b} + 2\bar{x}\,\mathrm{cov}(\bar{y}, \hat{b})$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 \times \frac{\sigma^2}{S_{xx}} =$$

$$(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2.$$

$$(4)\mathrm{cov}(\hat{a}, \hat{b}) = \mathrm{cov}(\bar{y} - \hat{b}\bar{x}, \hat{b}) = \mathrm{cov}(\bar{y}, \hat{b}) - \bar{x}\,\mathrm{cov}(\hat{b}, \hat{b})$$

$$= 0 - \bar{x}D\hat{b} = -\frac{\bar{x}\sigma^2}{S_{xx}}.$$

$$(5)EQ_e = E(S_{yy} - \hat{b}^2 S_{xx}) = E\sum_{i=1}^{n}(y_i - \bar{y})^2 - S_{xx}E\hat{b}^2$$

$$= E(\sum_{i=1}^{n} y_i{}^2 - n\bar{y}^2) - S_{xx}(D\hat{b} + E^2\hat{b})$$

$$= \sum_{i=1}^{n} Ey_i^2 - nE\bar{y}^2 - S_{xx}(\frac{\sigma^2}{S_{xx}} + b^2)$$

$$= \sum_{i=1}^{n}(Dy_i + E^2 y_i) - n(D\bar{y} + E^2\bar{y}) - \sigma^2 - b^2 S_{xx}$$

$$= \sum_{i=1}^{n}(\sigma^2 + (a + bx_i)^2) - n(\frac{\sigma^2}{n} + (a + b\bar{x})^2) - \sigma^2 - b^2 S_{xx}$$

$$= n\sigma^2 + na^2 + 2abn\bar{x} + b^2\sum_{i=1}^{n} x_i{}^2 - \sigma^2$$

$$-na^2 - 2abn\bar{x} - n\bar{x}^2 b^2 - \sigma^2 - b^2 S_{xx}$$

$$= (n-2)\sigma^2 + b^2(\sum_{i=1}^{n} x_i{}^2 - n\bar{x}^2) - b^2 S_{xx} = (n-2)\sigma^2.$$

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

2) $\hat{b}\sim N\left(b,\frac{\sigma^2}{S_{xx}}\right)$, $\hat{a}\sim N\left(a,\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right)$.

证：$\because \hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{S_{xx}}$，$y_i\sim N(a+bx_i,\sigma^2)$,

$\therefore \hat{b}$为相互独立正态分布y_1，y_2,…，y_n的线性组合，

$\therefore$由正态分布线性独立可加性知，$\hat{b}$服从正态分布，

由1）知$\hat{b}\sim N\left(b,\frac{\sigma^2}{S_{xx}}\right)$.

$\because \hat{a} = \bar{y} - \hat{b}\bar{x}, \bar{y} \sim N(a + b\bar{x}, \frac{\sigma^2}{n}), \bar{y}$与$\hat{b}$不相关，

即$\bar{y}$与$\hat{b}$独立，

$\therefore \hat{a}$为相互独立正态分布$\bar{y}$与$\hat{b}$的线性组合，

$\therefore \hat{a} \sim N(a, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$.

4. 线性假设的显著性检验

1）线性效果的检验

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

检验假设： $H_0: b = 0,$

$H_1: b \neq 0.$

使用t检验法来进行检验

$$\hat{b} \sim N(b, \sigma^2 / S_{xx}),$$

$$\frac{Q_e}{\sigma^2} \sim \chi^2(n-2).$$

并且$\hat{b}, Q_e$相互独立,

证明

当H_0为真时$b=0$,

$$\text{则}\hat{b}\sim N(0,\frac{\sigma^2}{S_{xx}})\Leftrightarrow \hat{b}\sqrt{S_{xx}}\Big/\sigma\sim N(0,1)\Leftrightarrow \frac{\hat{b}^2S_{xx}}{\sigma^2}\sim\chi^2(1)\text{）},$$

$$Q_e=S_{yy}-\hat{b}^2S_{xx}\Leftrightarrow \frac{Q_e}{\sigma^2}=\frac{S_{yy}}{\sigma^2}-\frac{\hat{b}^2S_{xx}}{\sigma^2},$$

$$y_i\sim N(a+bx_i,\sigma^2),\therefore \frac{(n-1)\,S^2}{\sigma^2}=\frac{(n-1)\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2}$$

$$=\frac{S_{yy}}{\sigma^2}\sim\chi^2(n-1)$$

$$\therefore \frac{Q_e}{\sigma^2}\sim\chi^2(n-2)\text{，即}E(\frac{Q_e}{\sigma^2})=n-2\Leftrightarrow E(\frac{Q_e}{n-2})=\sigma^2.$$

$$(\hat{b}-b)\Big/\sqrt{\sigma^2/S_{xx}}\Big/\sqrt{\frac{Q_e}{\sigma^2}\Big/(n-2)}=\frac{(\hat{b}-b)\sqrt{S_{xx}}}{\hat{\sigma}}\sim t(n-2),$$

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

当 H_0 为真时 $b=0$,此时

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

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;而存在其他关系.

(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$\hat{b}=0.48303,\ S_{xx}=8250,\quad \hat{\sigma}^2=0.90,$

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

假设$H_0: b=0$的拒绝域为

$$|t|=\frac{|\hat{b}|}{\hat{\sigma}}\sqrt{S_{xx}}\geq 2.3060,$$

$$|t|=\frac{0.48303}{\sqrt{0.90}}\times\sqrt{8250}=46.25,$$

拒绝 $H_0: b=0$, 认为回归效果显著.

2）线性系数b的检验

假设$H_0: b=k, H_1: b \neq k$,若H_0成立，

则$\hat{b} \sim N(k, \frac{\sigma^2}{S_{xx}}) \Leftrightarrow \frac{(\hat{b}-k)\sqrt{s_{xx}}}{\sigma} \sim N(0,1)$,而$\frac{Q_e}{\sigma^2} \sim \chi^2(n-2)$，

$$\therefore \text{取} T = \frac{\frac{(\hat{b}-k)\sqrt{s_{xx}}}{\sigma}}{\sqrt{\frac{Q_e}{\sigma^2(n-2)}}} = \frac{(\hat{b}-k)\sqrt{s_{xx}}}{\sqrt{\frac{Q_e}{n-2}}}$$

$$= \frac{(\hat{b}-k)\sqrt{s_{xx}}}{\hat{\sigma}} \sim t(n-2)$$

令$p\{|t| \geq t_{\frac{\alpha}{2}}(n-2)\} = \alpha$,计算$|T|$与$t_{\frac{\alpha}{2}}(n-2)$比较做出判断.

5. Y 的观察值的点预测和预测区间

若拒绝H_0,则表明随机变量y与x线性相关关系显著，此时，当$x = x_0$，则利用线性回归方程可预测对应y的值.

1、点预测(估计)：

$\hat{y}_0 = \hat{a} + \hat{b}x_0$，且$\hat{y}_0$为$y_0$的无偏估计.

$\because E\hat{y}_0 = E(\hat{a} + \hat{b}x_0) = a + bx_0 = Ey_0$,

$y_0 \sim N(a + bx_0, \sigma^2)$.

2、区间预测(估计)：

$\because \bar{y}$与$\hat{b}$服从正态分布，$\mathrm{cov}(\bar{y},\hat{b})=0$,

$\therefore \bar{y}$与$\hat{b}$相互独立.

又$\because \hat{y}_0=\hat{a}+\hat{b}x_0=\bar{y}+\hat{b}(x_0-\bar{x})$,

$$\therefore D\hat{y}_0=D\bar{y}+(x_0-\bar{x})^2D\hat{b}=\sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right).$$

$$\therefore \hat{y}_0\sim N\left(a+bx_0,\sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right),$$

或$D\hat{y}_0=D(\hat{a}+\hat{b}x_0)$

$=D\hat{a}+x_0^2D\hat{b}+2x_0\,\mathrm{cov}(\hat{a},\hat{b})$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) + \frac{x_0^2\sigma^2}{S_{xx}} - \frac{2x_0\bar{x}\sigma^2}{S_{xx}}$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}(x_0^2 - 2x_0\bar{x} + \bar{x}^2) = \frac{\sigma^2}{n} + \frac{\sigma^2(x_0 - \bar{x})^2}{S_{xx}}$$

$$= \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right). \therefore \hat{y}_0 \sim N\left(a + bx_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right),$$

$\because \hat{y}_0$依赖于$y_1, y_2, \cdots, y_n$，$y_0 \sim N(a + bx_0, \sigma^2)$，$y_0$与$\hat{y}_0$相互独立，

$$\therefore D(y_0 - \hat{y}_0) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right), E(y_0 - \hat{y}_0) = 0,$$

$$\therefore y_0 - \hat{y}_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right).$$

设 Y_0 是在 $x = x_0$ 处对 Y 的观察结果.

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

x_0处的经验回归函数值 $\hat{Y}_0 = \hat{a} + \hat{b}x_0$

作为$Y_0 = a + bx_0 + \varepsilon_0$的点预测

Y_0是将要做一次独立试验的结果，

它与已经得到的试验结果$Y_1, Y_2, \cdots, Y_n$相互独立.

$\hat{b}$是$Y_1, Y_2, \cdots, Y_n$的线性组合，

故$\hat{Y}_0 = \bar{Y} + \hat{b}(x_0 - \bar{x})$是$Y_1, Y_2, \cdots, Y_n$的线性组合，

Y_0与$\hat{Y}_0$相互独立

$$\hat{Y}_0 - Y_0 \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\sigma^2\right),$$

$$\frac{\hat{Y}_0 - Y_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0,1)$$

$Y_0, \hat{Y}_0, Q_e$的相互独立性知

$$\frac{\hat{Y}_0 - Y_0}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}} \Big/ \sqrt{\frac{Q_e}{\sigma^2}\Big/(n-2)} =$$

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim t(n-2), \text{给定置信水平为} 1-\alpha,$$

$$P\left\{\left|\hat{Y}_0 - Y_0\right| \Big/ \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} \le t_{\alpha/2}(n-2)\right\}$$

$$= 1-\alpha$$

$$P\left\{\hat{Y}_0 - t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} < Y_0\right.$$

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

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

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

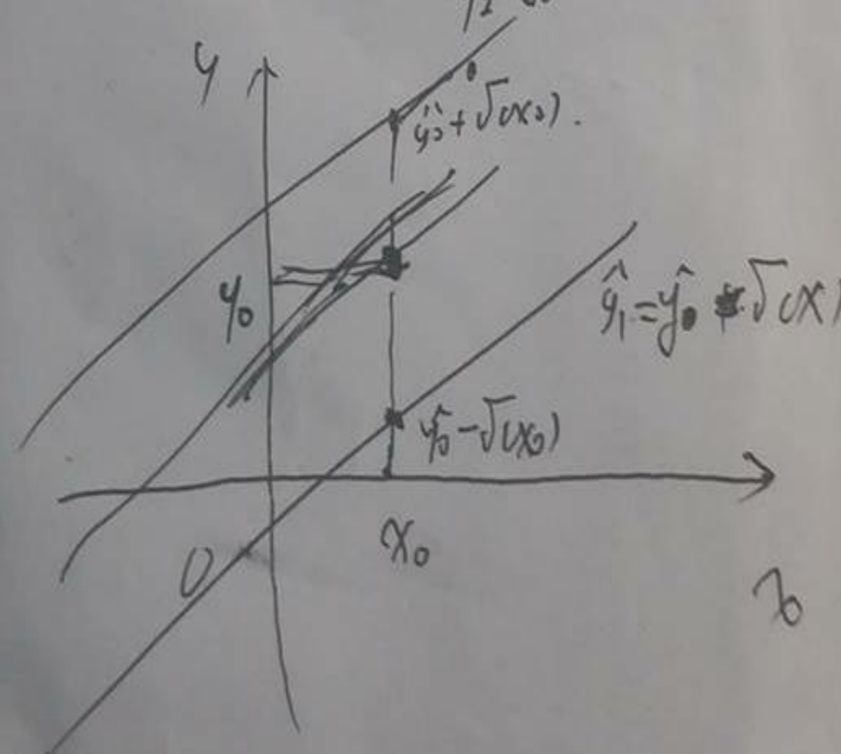
Y_0 的置信水平为$1-\alpha$的预测区间

$$\text{记}\delta(x_0)=t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}},$$

Y_0 的置信水平为$1-\alpha$ 的预测区间

$$\left(y_1=\hat{Y}_0-\delta(x_0),\quad \hat{Y}_0+\delta(x_0)=y_2\right)$$

※ 1° 给 x_0 的值，置信度 $1-\alpha$，则 y 的预测区间为 $[\hat{y}_0-\delta(x_0),\ \hat{y}_0+\delta(x_0)]$，如图所示



2°. 当 n 很大时，且 x 在 $\bar{x}$ 附近取值，则

$$t_{\frac{\alpha}{2}}(n-2)\approx u_{\frac{\alpha}{2}},\quad \lim_{n\to\infty}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{L_{xx}}}=1.$$

此时：$\delta(x)\approx\hat{\sigma}_e u_{\frac{\alpha}{2}}$.

则 y 的预测区间为 $[\hat{y}-\hat{\sigma}u_{\frac{\alpha}{2}},\ \hat{y}+\hat{\sigma}u_{\frac{\alpha}{2}}]$.

3° 当 x 在 $\bar{x}$ 附近，置信区间越窄，则预测精度越高。$\hat{\sigma}$ 越小，预测精度越高

例5（续例2） 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.**

解 $\hat{Y}_0=\hat{Y}\big|_{x=125}=[-2.73935+0.48303x]_{x=125}=57.64,$

在$x=125$处Y的新观察值Y_0的一个置信水平为**0.95的预测区间为**

$$\left(\hat{Y}_0\Big|_{x=x_0}\pm t_{0.025}(8)\hat{\sigma}\sqrt{1+\frac{1}{10}+\frac{(x_0-145)^2}{8250}}\right)$$

6. 控制

考虑 $y=\beta_0+\beta_1 x+\varepsilon$ 的回归方程 若使置信度为 $1-\alpha$ 的区间 $[y_1, y_2]$ 内，那么应该

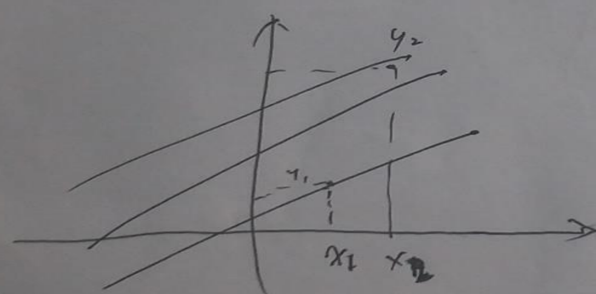
给 x 应控制在什么范围。

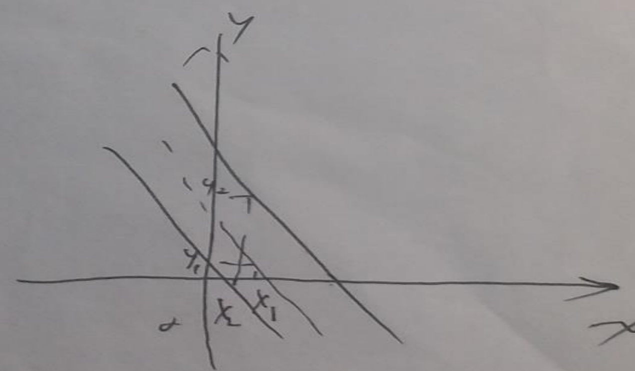
由预测区间推导，当 n 很大时

有
$$\begin{cases} y_1=\hat{y}_1-u_{\frac{\alpha}{2}}\hat{\sigma}_e=\hat{\beta}_0+\hat{\beta}_1 x_1-u_{\frac{\alpha}{2}}\hat{\sigma} \\ y_2=\hat{y}_2+u_{\frac{\alpha}{2}}\hat{\sigma}_e=\hat{\beta}_0+\hat{\beta}_1 x_2+u_{\frac{\alpha}{2}}\hat{\sigma} \end{cases}$$

$$\Longleftrightarrow \begin{cases} x_1=\frac{1}{\hat{\beta}_1}(y_1+\hat{\sigma}u_{\frac{\alpha}{2}}-\hat{\beta}_0) \\ x_2=\frac{1}{\hat{\beta}_1}(y_2-\hat{\sigma}u_{\frac{\alpha}{2}}-\hat{\beta}_0) \end{cases}$$

若 $\hat{\beta}_1>0$，x 的控制区间为 (x_1, x_2)。若 $\hat{\beta}_1<0$，控制 x 的区间为 $[x_2, x_1]$





[注：线性回归问题 ~~[illegible]~~ 也可以推广到多元线性回归问题及可化为

多元线性回归问题分析。可用同样方法，对非线性回归 ……

三、可化为一元线性回归的例子

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.

几种常见的可转化为一元线性回归的模型

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

其中α,β,σ^2是与x无关的未知参数.

将$Y=\alpha e^{\beta x}\cdot\varepsilon$两边取对数,

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

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

转化为一元线性回归模型：

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

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

其中α,β,σ^2是与x无关的未知参数.

将$Y=\alpha e^{\beta x}\cdot\varepsilon$两边取对数，

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

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

转化为一元线性回归模型：

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

3. $Y=\alpha+\beta h(x)+\varepsilon, \qquad \mathcal{E}\sim N(0,\sigma^2).$

其中α,β,σ^2是与x无关的未知参数.

$h(x)$是x的已知函数，

令$\alpha=a, \quad \beta=b, \quad h(x)=x'$,

转化为一元线性回归模型：

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

例6 下表是 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 .

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

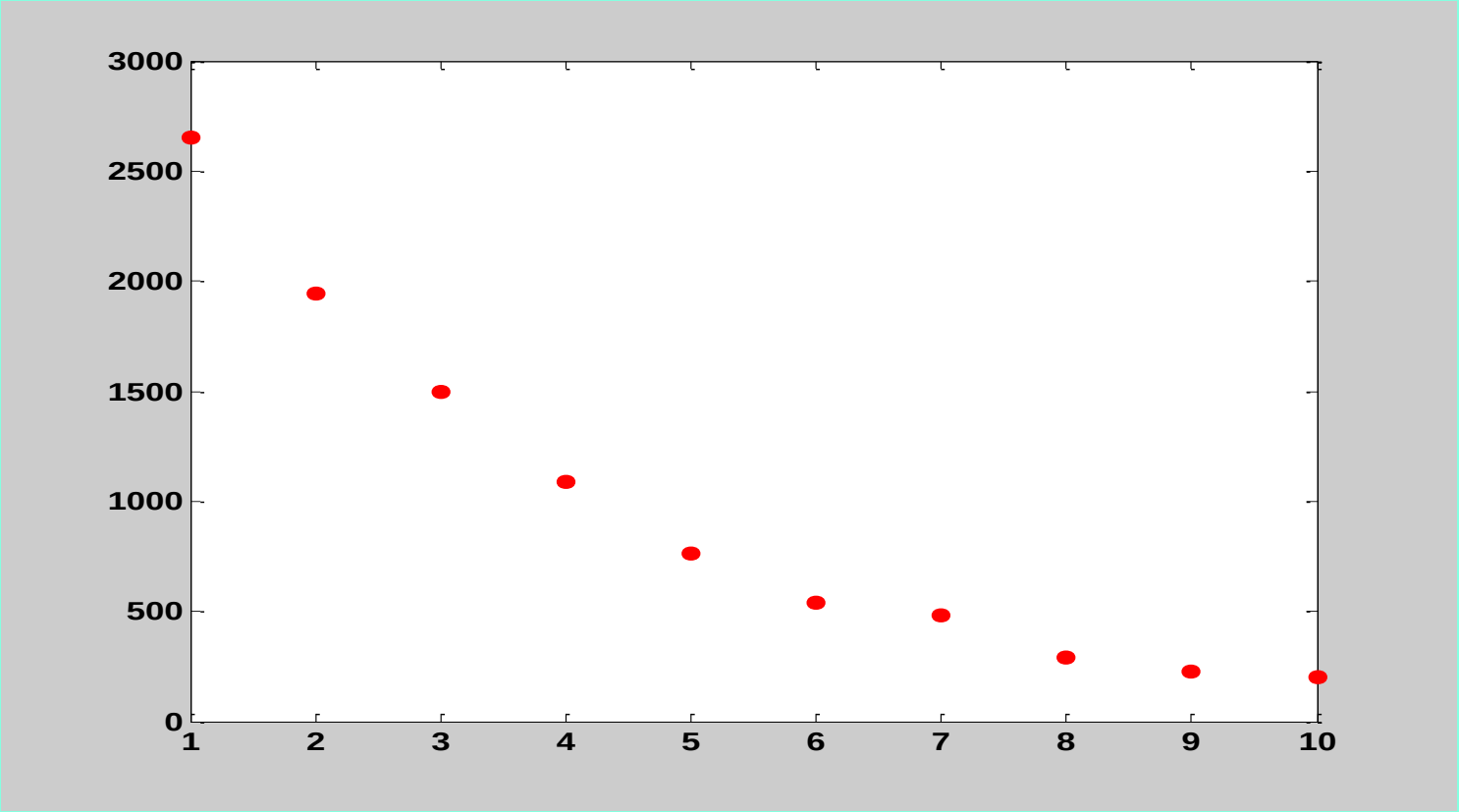
解

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')
```

Y与x呈指数关系，



选择模型 $Y=\alpha e^{\beta x}\cdot\varepsilon,\qquad \ln\varepsilon\sim N(0,\sigma^2).$

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

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

数据变换后得

$x'=x$	1	2	3	4	5
$y'=\ln y$	7.8827	7.5720	7.3092	6.9912	6.6399

$x'=x$	6	7	8	9	10
$y'=\ln y$	6.2879	6.1821	5.6699	5.4205	5.3181

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

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

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、多元线性回归的数学模型

1、数学模型

实际问题中的随机变量Y 通常与多个普通变量$x_1, x_2, \cdots, x_p$ $(p>1)$ 有关.

对于自变量 $x_1, x_2, \cdots, x_p$ 的一组确定值, Y 具有一定的分布,

假设 $Y = b_0 + b_1 x_1 + \cdots + b_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$

$b_0, b_1, \cdots, b_p, \sigma^2$是与$x_1, \cdots, x_p$无关的未知参数.

则理论回归方程为$Y' = b_0 + b_1 x_1 + \cdots + b_p x_p.$

2、数学模型的分析与求解

设 $(x_{11}, x_{12},\cdots, x_{1p}, y_1),\cdots,(x_{n1}, x_{n2},\cdots, x_{np}, y_n)$ 是一个样本.

用最大似然估计法估计参数.

取 $\hat{b}_0,\hat{b}_1,\cdots,\hat{b}_p$，当 $b_0=\hat{b}_0,\ b_1=\hat{b}_1,\cdots,b_p=\hat{b}_p$ 时，

$$Q=\sum_{i=1}^{n}(y_i-b_0-b_1x_{i1}-\cdots-b_px_{ip})^2$$

达到最小.

求Q分别关于$b_0, b_1, \cdots, b_p$的偏导数，

并令它们等于零，得

$$\left.\begin{aligned} \frac{\partial Q}{\partial b_0} &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip}) = 0, \\ \frac{\partial Q}{\partial b_j} &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip}) x_{ij} = 0, \\ & \qquad j = 1, 2, \cdots, p \end{aligned}\right\}$$

化简可得

$$
\begin{cases}
b_0 n + b_1 \sum_{i=1}^{n} x_{i1} + b_2 \sum_{i=1}^{n} x_{i2} + \cdots + b_p \sum_{i=1}^{n} x_{ip} = \sum_{i=1}^{n} y_i, \\
b_0 \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + b_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + b_p \sum_{i=1}^{n} x_{i1} x_{ip} \\
\qquad = \sum_{i=1}^{n} x_{i1} y_i, \\
\cdots \\
b_0 \sum_{i=1}^{n} x_{ip} + b_1 \sum_{i=1}^{n} x_{ip} x_{i1} + b_2 \sum_{i=1}^{n} x_{ip} x_{i2} + \cdots + b_p \sum_{i=1}^{n} x_{ip}^2 \\
\qquad = \sum_{i=1}^{n} x_{ip} y_i .
\end{cases}
$$

正规方程组

引入矩阵

$$X=\begin{pmatrix}1 & x_{11} & x_{12} & \cdots & x_{1p}\\ 1 & x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{np}\end{pmatrix},\quad Y=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix},\quad B=\begin{pmatrix}b_0\\ b_1\\ \vdots\\ b_p\end{pmatrix}.$$

$$X^{\mathrm{T}}X=\begin{pmatrix}1 & 1 & \cdots & 1\\ x_{11} & x_{21} & \cdots & x_{n1}\\ \vdots & \vdots & & \vdots\\ x_{1p} & x_{2p} & \cdots & x_{np}\end{pmatrix}\begin{pmatrix}1 & x_{11} & x_{12} & \cdots & x_{1p}\\ 1 & x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{np}\end{pmatrix},$$

$$= \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ip} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \cdots & \sum_{i=1}^{n} x_{i1} x_{ip} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ip} & \sum_{i=1}^{n} x_{ip} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ip}^2 \end{pmatrix}$$

$$X^{\mathrm{T}}Y=\begin{pmatrix}1 & 1 & \cdots & 1\\ x_{11} & x_{21} & \cdots & x_{n1}\\ \vdots & \vdots & & \vdots\\ x_{1p} & x_{2p} & \cdots & x_{np}\end{pmatrix}\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}=\begin{pmatrix}\sum_{i=1}^{n} y_i\\ \sum_{i=1}^{n} x_{i1}y_i\\ \vdots\\ \sum_{i=1}^{n} x_{ip}y_i\end{pmatrix}$$

正规方程组的矩阵形式

$$X^T XB = X^T Y$$

$$\hat{B}=\begin{pmatrix}\hat{b}_0\\ \hat{b}_1\\ \vdots\\ \hat{b}_p\end{pmatrix}=(X^TX)^{-1}X^TY$$ 最大似然估计值

取 $\hat{b}_0+\hat{b}_1x_1+\cdots+\hat{b}_px_p$ 记成 $\hat{y}$

作为$Y'(x_1,x_2,\cdots,x_p)=b_0+b_1x_1+\cdots+b_px_p$的估计,

方程$\hat{y}=\hat{b}_0+\hat{b}_1x_1+\hat{b}_2x_2+\cdots+\hat{b}_px_p$

称为*P*元经验线性回归方程, 简称回归方程.

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p-1},$$

补充逆矩阵计算

定义：设 A_{ij} 是矩阵

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

中元素 a_{ij} 的代数余子式，矩阵

$$A^{*}=\begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & A_{nm} \end{bmatrix}$$

称为 A 的伴随矩阵。

注意：伴随矩阵中的元素 A_{ij} 是按转置的顺序排列的。

则 $A^{-1}=\dfrac{1}{\Delta}A^{*}$ （$\Delta=|A|\neq 0$）.

在 n 阶行列式中，把元素 a_{ij} 所在的第 i 行和第 j 列划去后，留下来的 $n-1$ 阶行列式叫做元素 a_{ij} 的余子式，记作 M_{ij}.

记 $A_{ij}=(-1)^{i+j}M_{ij}$，叫做元素 a_{ij} 的代数余子式.

例如 $D=\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}$

$M_{23}=\begin{vmatrix} a_{11} & a_{12} & a_{14} \\ a_{31} & a_{32} & a_{34} \\ a_{41} & a_{42} & a_{44} \end{vmatrix}$

3、一元多项式回归

一元多项式回归可化为多元线性回归求解.

$$y' = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0$$

令$x_1 = x, x_2 = x^2, \cdots, \ x_m = x^m,$

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

其中$b_0, b_1, \cdots, b_m, \sigma^2$是与$x_1, \cdots, x_m$无关的未知参数.

例7　某件产品每件平均单价Y(元)与批量x(件)之间的关系的一组数据

x	20	25	30	35	40	50	55	60	65	70	80	90
y	1.81	1.70	1.65	1.55	1.48	1.40	1.30	1.26	1.24	1.21	1.20	1.18

解　$Y = b_0 + b_1 x + b_2 x^2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$

令$x_1 = x, \quad x_2 = x^2,$

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

$$X = \begin{pmatrix} 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 30 & 900 \\ 1 & 35 & 1225 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 65 & 4225 \\ 1 & 70 & 4900 \\ 1 & 75 & 5625 \\ 1 & 80 & 6400 \\ 1 & 90 & 8100 \end{pmatrix} \quad Y = \begin{pmatrix} 1.81 \\ 1.70 \\ 1.65 \\ 1.55 \\ 1.48 \\ 1.40 \\ 1.30 \\ 1.26 \\ 1.24 \\ 1.21 \\ 1.20 \\ 1.18 \end{pmatrix} \quad B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$$

$$X^{\mathrm{T}}X=\begin{pmatrix}12 & 640 & 40100\\ 640 & 40100 & 2779000\\ 40100 & 2779000 & 204702500\end{pmatrix}$$

$$(X^{\mathrm{T}}X)^{-1}=$$

$$\frac{1}{\Delta}\begin{pmatrix}4.8572925\times10^{11} & -1.95717\times10^{10} & 170550000\\ -1.95717\times10^{10} & 848420000 & -7684000\\ 170550000 & -7684000 & 71600\end{pmatrix}$$

$\Delta=1.41918\times10^{11}$,正规方程组的解为

$$\hat{B}=\begin{pmatrix}\hat{b}_0\\ \hat{b}_1\\ \hat{b}_2\end{pmatrix}=(X^TX)^{-1}X^TY=(X^TX)^{-1}\begin{pmatrix}16.98\\ 851.3\\ 51162\end{pmatrix}$$

$$=\begin{pmatrix}2.19826629\\ -0.02252236\\ 0.00012507\end{pmatrix}$$

得到回归方程

$$\hat{y}=2.19826629-0.02252236x+0.00012507x^2$$

逐步回归分析

在实际问题中,影响因变量的因素很多, 而这些因素之间可能存在多重共线性. 为得到可靠的回归模型, 需要一种方法能有效地从众多因素中挑选出对因变量贡献大的因素.

如果采用多元线性回归分析,回归方程稳定性差,每个自变量的区间误差积累将影响总体误差, 预测的可靠性差、精度低; 另外, 如果采用了影响小的变量,遗漏了重要变量,可能导致估计量产生偏倚和

不一致性.

“最优”的回归方程应该包含所有有影响的变量而不包括影响不显著的变量.

选择“最优”回归方程的方法

1.从所有可能的变量组合的回归方程中选择最优者;

2.从包含全部变量的回归方程中逐次剔除不显著因子;

3.从一个变量开始,把变量逐个引入方程;

4. “有进有出”的逐步回归分析.

逐步回归分析法在筛选变量方面比较理想,是目前较常用的方法.它从一个自变量开始,根据自变量作用的显著程度,从大到小地依次逐个引入回归方程,但当引入的自变量由于后面变量的引入而变得不显著时,要将其剔除掉.引入一个自变量或从回归方程中剔除一个自变量,为逐步回归的一步,对于每一步,都进行检验,以确保每次引入新的显著性变量前回归方程中只包含作用显著的变量.

练习题

在钢线碳含量对于电阻的效应的研究中，得到如下表所示一批数据：

含碳量 x %	0.1	0.3	0.40	0.55	0.70	0.80	0.95
电阻 $y(\Omega)$	15	18	19	21	22.6	23.8	26

(1)假设随机误差 ε 服从 $N(0,\sigma^2)$ 分布，试求 y 关于 x 的经验回归方程; (2)设 $\alpha = 0.05$,试问 y 关于 x 的线性回归方程效果是否显著？ (3)在 $x = 0.6$ 时，求出 y 的预测值及置信度为95%的预测区间.