Matlab概率统计

1. 1 统计工具箱中的基本统计命令
   - 基本统计量
   - 常见概率分布的函数
   - 参数估计
   - 假设检验
2. 2 回归分析
   - 1.多元线性回归
   - 2.多元二项式回归
   - 3.逐步回归分析
   - 4.非线性回归

基本统计量

对随机变量x，计算其基本统计量的命令如下：

- 均值： mean(x)
- 中位数： median(x)
- 标准差： std(x)
- 方差： var(x)
- 偏度： skewness(x) 峰度： kurtosis(x)

常见概率分布的函数

- 常见的几种分布的命令字符为：正态分布：norm; 指数分布：exp; 帕松分布：poiss; β 分布：beta; 威布尔分布：weib; 分布：chi2; t分布：t; F分布：F.
- Matlab工具箱对每一种分布都提供五类函数，其命令字符为：概率密度：pdf; 概率分布：cdf; 逆概率分布：inv; 均值与方差：stat; 随机数生成：rnd.
- 当需要一种分布的某一类函数时，将以上所列的分布命令字符与函数命令字符接起来，并输入自变量（可以是标量、数组或矩阵）和参数即可
- 如对均值为mu、标准差为sigma的正态分布，举例如后。

密度函数

p=normpdf(x,mu,sigma) (当mu=0,sigma=1时可缺省)

- 例画出正态分布N(0,1) 和N(0,4) 的概率密度函数图形.
- 在Matlab中输入以下命令:
  - x=-6:0.01:6;
  - y=normpdf(x);
  - z=normpdf(x,0,2); plot(x,y,x,z)

概率分布

P=normcdf(x,mu,sigma)

- 计算标准正态分布的概率$P\{-1 < X < 1\}$.
- 命令为：P=normcdf(1)-normcdf(-1)
- 结果为：P =0.6827

逆概率分布

x=norminv(P0,mu,sigma). 即求出x，使得$P\{X < x\} = P0$.此命令可用来求分位数.

- 取$alpha = 0.05$, 求$u_{1-\frac{\alpha}{2}}$
- $u_{1-\frac{\alpha}{2}}$ 的含义是: $X \sim N(0,1), P\{X < u_{1-\frac{\alpha}{2}}\} = 1 - \frac{\alpha}{2}$
- $\alpha = 0.05$,P=0.975
- $u_{0.975} = norminv(0.975) = 1.96$

均值与方差

[m,v]=normstat(mu,sigma)

- 求正态分布N(3,25)的均值与方差
- 命令为：[m,v]=normstat(3,5)

随机数生成

normrnd(mu,sigma,m,n).产生m*n阶的正态分布随机数矩阵.

- n2 = normrnd(0,1,[1 5])
- n2 = 0.0591 1.7971 0.2641 0.8717 -1.4462

频数直方图的描绘

- [N,X]=hist(data,k)
- 此命令将区间[min(data),max(data)]分为k个小区间（缺省为10），返回数组data落在每一个小区间的频数N和每一个小区间的中点X.
- 绘制数组data的频数直方图需要这样调用： hist(data,k)

1、正态总体的参数估计

设总体服从正态分布，则其点估计和区间估计可同时由以下命令获得:

- [muhat,sigmahat,muci,sigmaci] = normfit(X,alpha)
- 此命令在显著性水平alpha下估计数据X的参数（alpha缺省时设定为0.05）
- 返回值muhat是X的均值的点估计值
- muci是均值的区间估计
- sigmahat是标准差的点估计值
- sigmaci是标准差的区间估计

2、其它分布的参数估计

有两种处理办法:

- 一.取容量充分大的样本（$n > 50$），按中心极限定理，它近似地服从正态分布;
- 二.使用Matlab工具箱中具有特定分布总体的估计命令.
  - （1）[muhat, muci] = expfit(X,alpha)—— 在显著性水平alpha下，求指数分布的数据X的均值的点估计及其区间估计.
  - （2）[lambdahat, lambdaci] = poissfit(X,alpha)—— 在显著性水平alpha下，求泊松分布的数据X 的参数的点估计及其区间估计.
  - [phat, pci] = weibfit(X,alpha)—— 在显著性水平alpha下，求Weibull分布的数据X 的参数的点估计及其区间估计.

假设检验

在总体服从正态分布的情况下，可用以下命令进行假设检验.

- 1、总体方差$sigma^2$已知时，总体均值的检验使用z-检验
- [h,sig,ci] = ztest(x,m,sigma,alpha,tail)
- 检验数据x 的关于均值的某一假设是否成立，
- 其中sigma 为已知方差，
- alpha 为显著性水平

总体方差$sigma^2$已知时，总体均值的检验使用z-检验

究竟检验什么假设取决于tail 的取值:

- tail = 0，检验假设 “x 的均值等于m ”
- tail = 1，检验假设 “x 的均值大于m ”
- tail =-1，检验假设 “x 的均值小于m ”
- tail的缺省值为0，alpha的缺省值为0.05.

总体方差$sigma^2$已知时，总体均值的检验使用z-检验

- 返回值h 为一个布尔值，h=1 表示可以拒绝假设，
- h=0 表示不可以拒绝假设，
- sig 为假设成立的概率，
- ci 为均值的1-alpha 置信区间. .

Matlab实验

- 例: Matlab统计工具箱中的数据文件gas.mat.中提供了美国1993年一月份和二月份的汽油平均价格（price1,price2分别是一，二月份的油价，单位为美分），它是容量为20的双样本.假设一月份油价的标准偏差是一加仑四分币（$\sigma = 4$），试检验一月份油价的均值是否等于115.
- load gas
- [h,sig,ci] = ztest(price1,115,4)

检验结果

- 返回： h = 0， sig = 0.8668， ci = [113.3970 116.9030].
- 1. 布尔变量h=0, 表示不拒绝零假设. 说明提出的假设均值115 是合理的.
- 2. sig-值为0.8668, 远超过0.5, 不能拒绝零假设
- 3. 95%的置信区间为[113.4, 116.9], 它完全包括115, 且精度很高.

2、总体方差sigma2未知时，总体均值的检验使用t-检验

[h,sig,ci] = ttest(x,m,alpha,tail) 检验数据x 的关于均值的某一假设是否成立

- alpha 为显著性水平，究竟检验什么假设取决于tail 的取值:
- tail = 0，检验假设 “x 的均值等于m ”
- tail = 1，检验假设 “x 的均值大于m ”
- tail =-1，检验假设 “x 的均值小于m ”
- tail的缺省值为0，alpha的缺省值为0.05.

总体方差sigma2未知时，总体均值的检验使用t-检验

- 返回值h 为一个布尔值，h=1 表示可以拒绝假设，
- h=0 表示不可以拒绝假设，
- sig 为假设成立的概率，
- ci 为均值的1-alpha 置信区间

Matlab实验

- 例: 试检验例8中二月份油价Price2的均值是否等于115.
- 作假设: m = 115，price2为二月份的油价，不知其方差，故用以下命令检验
- [h,sig,ci] = ttest(price2 ,115)

检验结果

- 返回： h = 1， sig = 4.9517e-004， ci =[116.8 120.2].
- 1. 布尔变量h=1, 表示拒绝零假设. 说明提出的假设油价均值115是不合理的.
- 2. 95%的置信区间为[116.8 120.2], 它不包括115, 故不能接受假设.
- 3. sig-值为4.9517e-004, 远小于0.5, 不能接受零假设.

3、两总体均值的假设检验使用t-检验

[h,sig,ci] = ttest2(x,y,alpha,tail) 检验数据x，y 的关于均值的某一假设是否成立，其中alpha 为显著性水平

- 究竟检验什么假设取决于tail 的取值:
- tail = 0，检验假设 “x 的均值等于y 的均值”
- tail = 1，检验假设 “x 的均值大于y 的均值”
- tail =-1，检验假设 “x 的均值小于y 的均值”
- tail的缺省值为0，alpha的缺省值为0.05.

两总体均值的假设检验使用t-检验

- 返回值h 为一个布尔值，h=1 表示可以拒绝假设，
- h=0 表示不可以拒绝假设，
- sig 为假设成立的概率，
- ci 为x与y均值差的的1-alpha 置信区间.

Matlab实验

- 例：试检验前面例中一月份油价Price1与二月份的油价Price2均值是否相同.
- [h,sig,ci] = ttest2(price1,price2)
- 返回：h = 1，sig = 0.0083，ci =[-5.8,-0.9].
- 1. 布尔变量h=1, 表示拒绝零假设. 说明提出的假设“油价均值相同”是不合理的.
- 2. 95%的置信区间为[-5.8,-0.9],说明一月份油价比二月份油价约低1至6分.
- 3. sig-值为0.0083, 远小于0.5, 不能接受“油价均相同”假设

4、非参数检验：总体分布的检验

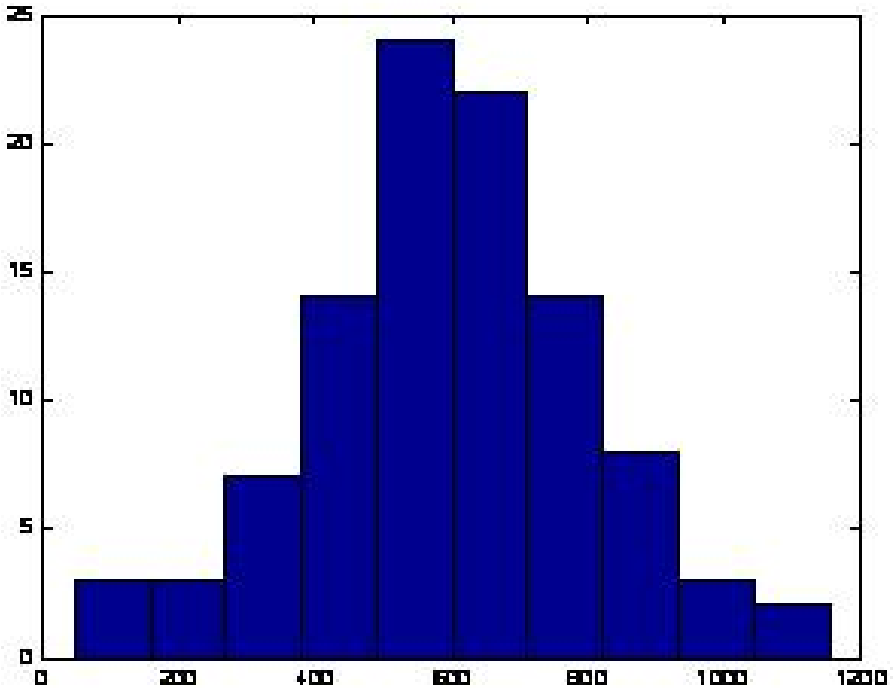
Matlab工具箱提供了两个对总体分布进行检验的命令:

- （1）h = normplot(x)
- 此命令显示数据矩阵x的正态概率图.如果数据来自于正态分布，则图形显示出直线性形态.而其它概率分布函数显示出曲线形态.
- （2）h = weibplot(x)
- 此命令显示数据矩阵x的Weibull概率图.如果数据来自于Weibull分布，则图形将显示出直线性形态.而其它概率分布函数将显示出曲线形态.

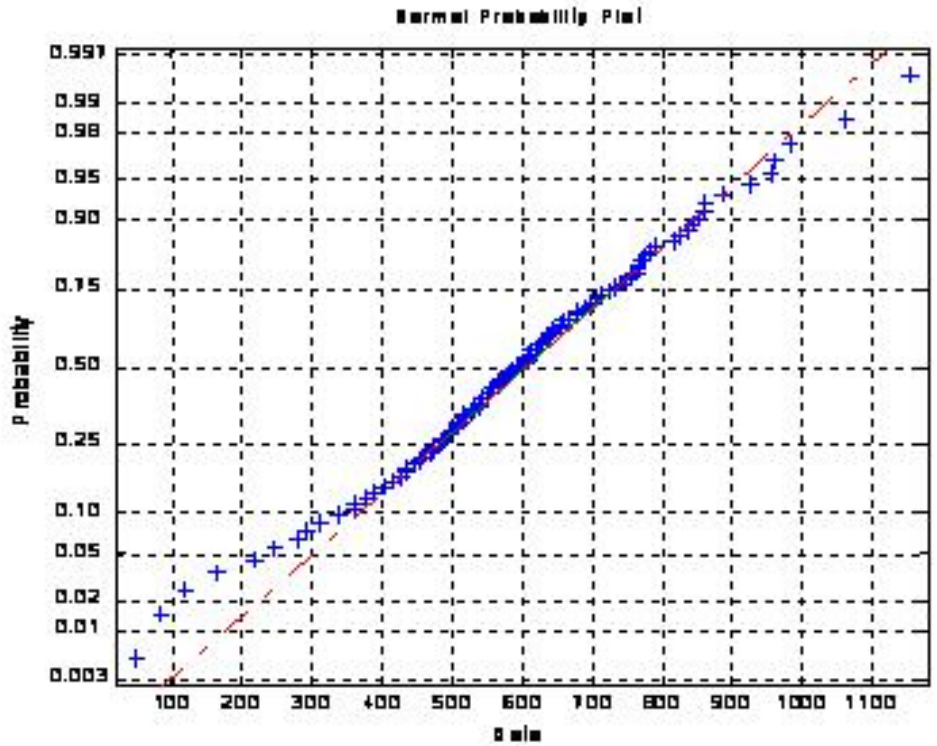
Matlab实验

- 一道工序用自动化车床连续加工某种零件，由于刀具损坏等会出现故障. 故障是完全随机的，并假定生产任一零件时出现故障机会均相同. 工作人员是通过检查零件来确定工序是否出现故障的.现积累有100次故障纪录，故障出现时该刀具完成的零件数如下:
- 459 362 624 542 509 584 433 748 815 505 612 452 434 982 640 742 565 706 593 680 926 653 164 487 734 608 428 1153 593 844 527 552 513 781 474 388 824 538 862 659 775 859 755 49 697 515 628 954 771 609 402 960 885 610 292 837 473 677 358 638 699 634 555 570 84 416 606 1062 484 120 447 654 564 339 280 246 687 539 790 581 621 724 531 512 577 496 468 499 544 645 764 558 378 765 666 763 217 715 310 851
- 试观察该刀具出现故障时完成的零件数属于哪种分布.

数据输入到x,作频数直方图hist(x,10)



分布的正态性检验：normplot(x)



Matlab实验

- 刀具寿命近似服从正态分布
- 参数估计：[muhat,sigmahat,muci,sigmaci] = normfit(x)
- 估计出该刀具的均值为594，方差204，
- 均值的0.95置信区间为[553.4962，634.5038]，
- 方差的0.95置信区间为[179.2276，237.1329].

Matlab实验

- 已知刀具的寿命服从正态分布，现在方差未知的情况下，检验其均值m 是否等于594.
- 结果：h = 0，sig = 1，ci =[553.4962，634.5038].
- 1. 布尔变量h=0, 表示不拒绝零假设. 说明提出的假设寿命均值594是合理的.
- 2. 95%的置信区间为[553.5，634.5], 它完全包括594, 且精度很高.
- 3. sig-值为1, 远超过0.5, 不能拒绝零假设.

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;
- 画出散点图，根据散点图确定曲线的类型;
  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

多元线性回归

- $$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$
- 1、确定回归系数的点估计值:
- b=regress(Y, X)
- $$b = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \ldots \\ \hat{\beta}_p \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$
- 对一元线性回归，取p=1即可

2、求回归系数的点估计和区间估计、并检验回归模型：

[b, bint,r,rint,stats]=regress(Y,X,alpha)

- alpha: 显著性水平（缺省时为0.05）; b 回归系数; bint:回归系数的区间估计
- r残差，rint 置信区间
- stats:用于检验回归模型的统计量，有三个数值：相关系数r^2、F值、与F对应的概率p
- 相关系数r^2越接近1，说明回归方程越显著;
- $F > F^{1-\sigma}(kn-k-1)$时拒绝H0，F越大，说明回归方程越显著;
- 与F对应的概率$p < \alpha$时拒绝H0，回归模型成立.

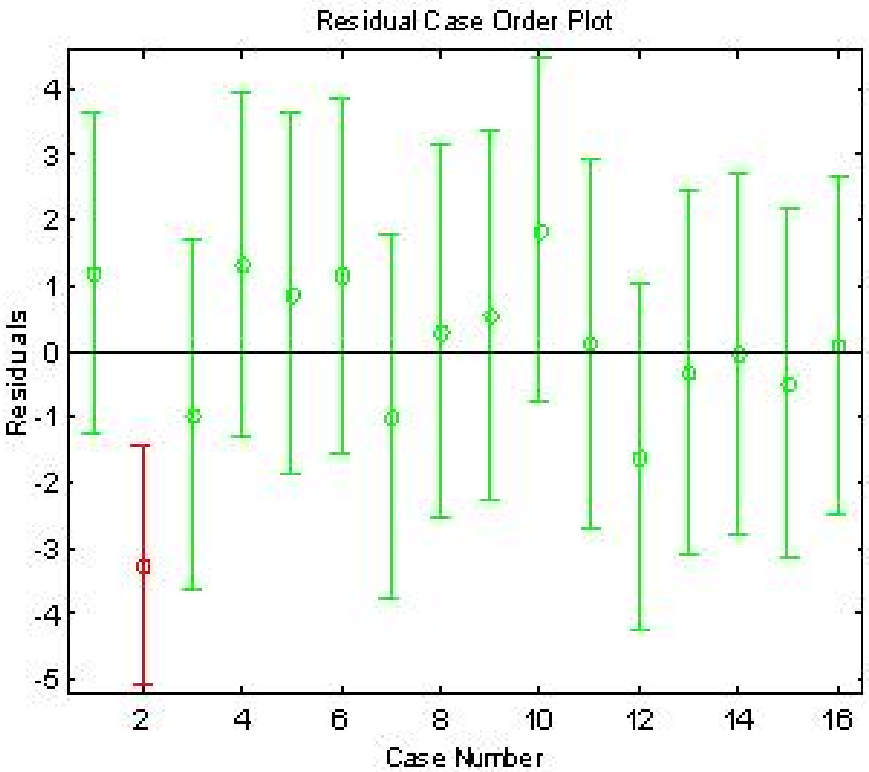
多元线性回归

- 3、画出残差及其置信区间：rcoplot（r，rint）
- 例1 输入数据：x=[143 145 146 147 149 150 153 154 155 156 157 158 159 160 162 164]';
- X=[ones(16,1) x];
- Y=[88 85 88 91 92 93 93 95 96 98 97 96 98 99 100 102]';
- [b,bint,r,rint,stats]=regress(Y,X)

结果

- b = -16.0730，0.7194
  bint = -33.7071 1.5612 0.6047 0.8340
  stats = 0.9282 180.9531 0.0000
- 即$\hat{\beta}_0 = -16.073, \hat{\beta}_1 = 0.7194$;
- $\hat{\beta}_0$ 的置信区间为[-33.7017，1.5612],
- $\hat{\beta}_1$ 的置信区间为[0.6047,0.834];
- r^2=0.9282, F=180.9531, p=0.0000 $p < 0.05$, 可知回归模型y=-16.073+0.7194x 成立

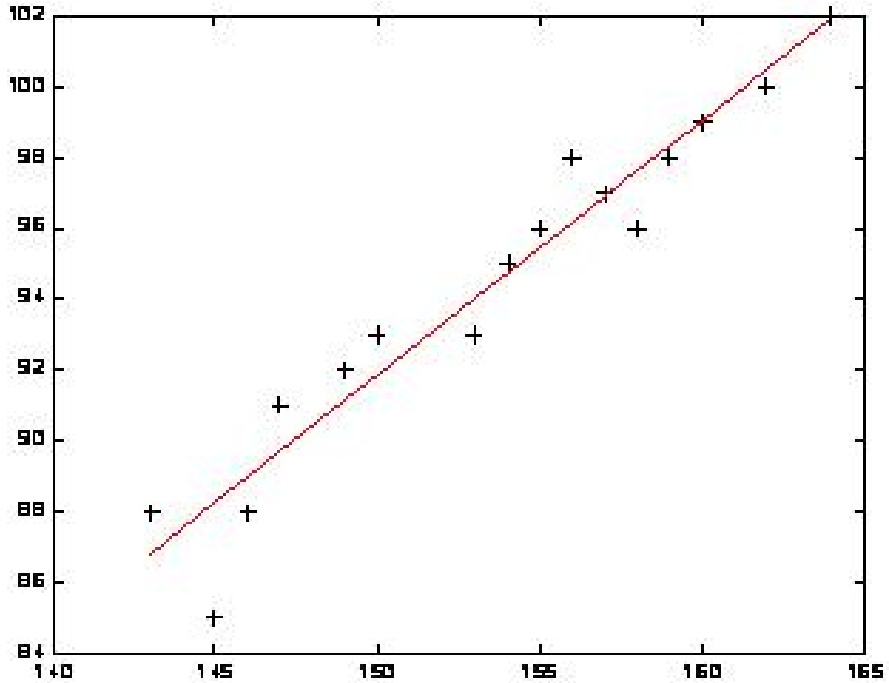
残差分析，作残差图：rcoplot(r,rint)



残差分析

- 从残差图可以看出，除第二个数据外，其余数据的残差离零点均较近，且残差的置信区间均包含零点
- 这说明回归模型y=-16.073+0.7194x能较好的符合原始数据，而第二个数据可视为异常点.

预测作图: z=b(1)+b(2)*x; plot(x,Y,'k+',x,z,'r')



多元二项式回归

- 命令：rstool（x，y，’model’，alpha）
- x,n*m矩阵
- y,n维列向量
- alpha,显著性水平（缺省时为0.05）

多元二项式回归

model可以是一下参数

- linear（线性）

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

- purequadratic（纯二次）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^{n} \beta_{jj} x_j^2$
- interaction（交叉）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \le j \ne k \le m} \beta_{jk} x_j x_k$
- quadratic（完全二次）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \le j,k \le m} \beta_{jk} x_j x_k$

Matlab实验

- x1=[1000 600 1200 500 300 400 1300 1100 1300 300];
- x2=[5 7 6 6 8 7 5 4 3 9];
- y=[100 75 80 70 50 65 90 100 110 60]';
- x=[x1' x2'];
- rstool(x,y,'purequadratic')

逐步回归分析法的思想

- 从一个自变量开始，视自变量Y作用的显著程度，从大到地依次逐个引入回归方程。
- 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。
- 对于每一步都要进行Y值检验，以确保每次引入新的显著性变量前回归方程中只包含对Y作用显著的变量。
- 这个过程反复进行，直至既无不显著的变量从回归方程中剔除，又无显著变量可引入回归方程时为止。

逐步回归的命令

- 逐步回归的命令
- stepwise（x，y，inmodel，alpha）
- x,自变量数据, m*n 阶矩阵
- 因变量数据，n维向量
- inmodel,矩阵的列数的指标，给出初始模型中包括的子集（缺省时设定为全部自变量）
- alpha,显著性水平（缺省时为0.5）

Matlab实验

- 例6 水泥凝固时放出的热量y与水泥中4种化学成分x1、x2、x3、x4 有关，今测得一组数据如下，试用逐步回归法确定一个线性模型.

- x1 = 7 1 11 11 7 11 3 1 2 21 1 11 10; x2= 26 29 56 31 52 55 71 31 54 47 40 66 68; x3= 6 15 8 8 6 9 17 22 18 4 23 9 8; x4= 60 52 20 47 33 22 6 44 22 26 34 12 12; y= 78.5 74.3 104.3 87.6 95.9 109.2 102.7 72.5 93.1 115.9 83.8 113.3 109.4

Matlab代码

- x1=[7 1 11 11 7 11 3 1 2 21 1 11 10]';
- x2=[26 29 56 31 52 55 71 31 54 47 40 66 68]';
- x3=[6 15 8 8 6 9 17 22 18 4 23 9 8]';
- x4=[60 52 20 47 33 22 6 44 22 26 34 12 12]';
- y=[78.5 74.3 104.3 87.6 95.9 109.2 102.7 72.5 93.1 115.9 83.8 113.3 109.4]';
- x=[x1 x2 x3 x4];

运行stepwise

- stepwise(x,y)
- 点击next,最后export,导出我们要的结果

非线性回归

三个命令:

- [beta，r，J]=nlinfit（x，y，’model’，beta0）
- nlintool（x，y，’model’，beta0，alpha）
- [Y，DELTA]=nlpredci（’model’，x，beta，r，J）
- 到第十章时候结合具体的例子讲