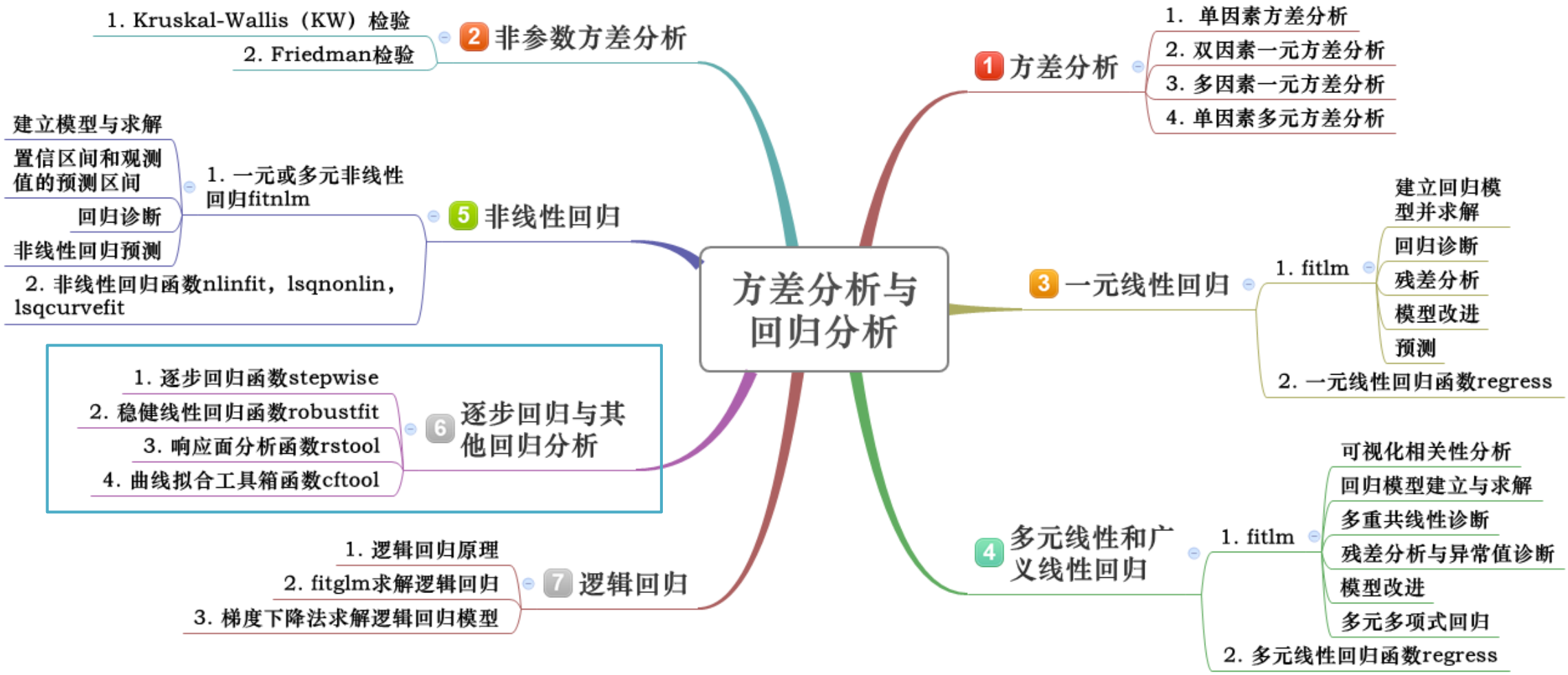
信陽師範学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

第11章 方差分析与回归分析

讲授人：牛言涛

日期：2020年4月17日

第11章 方差分析与回归分析知识点思维导图



一. 逐步回归简介

- “最优”的回归方程就是包含所有对Y有影响的变量，而不包含对Y影响不显著的变量回归方程。
- 选择“最优”的回归方程有以下几种方法：

  （1）从所有可能的因子（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。

- 以第四种方法，即逐步回归分析法在筛选变量方面较为理想.

一. 逐步回归简介

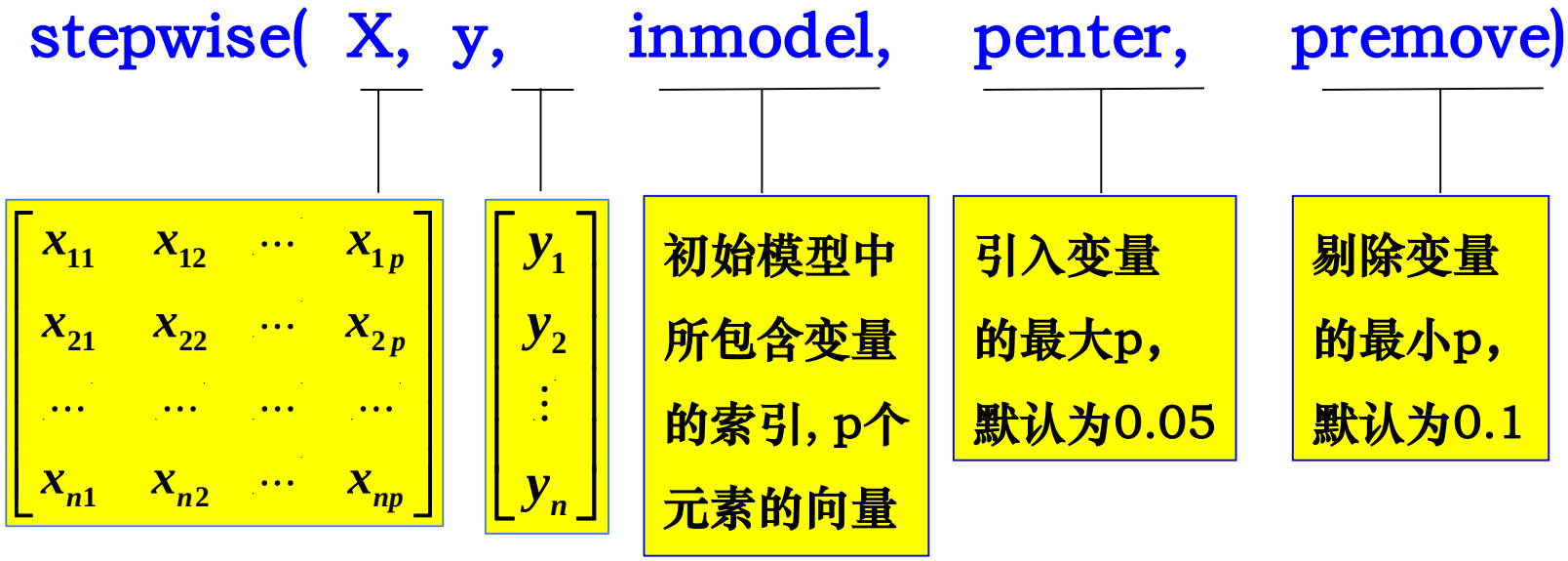
信阳师范学院
数学与统计学院

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。

1. 逐步回归函数stepwise

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS



指定了F统计的p值的初始模型（inmodel）和入口（penter）和出口（premove）容忍度（显著性水平），penter的值必须小于或等于premove的值。

函数运行后出现一交互式界面，通过该界面进行引入和剔除变量的操作，还可以导出相关结果。

stepwisefit针对多元线性模型，stepwiseglm针对广义线性模型逐步回归。

案例分析：光合作用

例：研究光合速率y与比叶重x1、气孔密度x2、叶绿素含量x3之间的关系，试验得到红薯性状观测值的数据如下表，试建立y关于x1，x2，x3的回归模型。

x1	x2	x3	y	x1	x2	x3	y
1.9993	11.4	4.0575	11.7161	1.9843	8.3	4.2719	9.8014
2.0254	8.1	3.7750	6.9862	1.9904	10.8	4.9872	11.0765
2.0010	10.7	3.3733	11.3444	1.7836	10.7	3.0019	6.3744
2.1072	11.2	3.1352	12.4770	1.9730	8.8	4.3073	9.3993
1.8941	9.0	3.5190	5.9618	1.9414	10.2	4.3965	9.8420
2.0188	12.5	3.4278	11.2210	2.0519	9.0	4.1673	8.2510
1.9362	10.1	3.8518	8.8416	1.9626	11.1	4.0186	10.6400
2.1072	8.5	4.1373	7.9488	1.8651	14.2	3.4175	6.6433

逐步回归建立模型

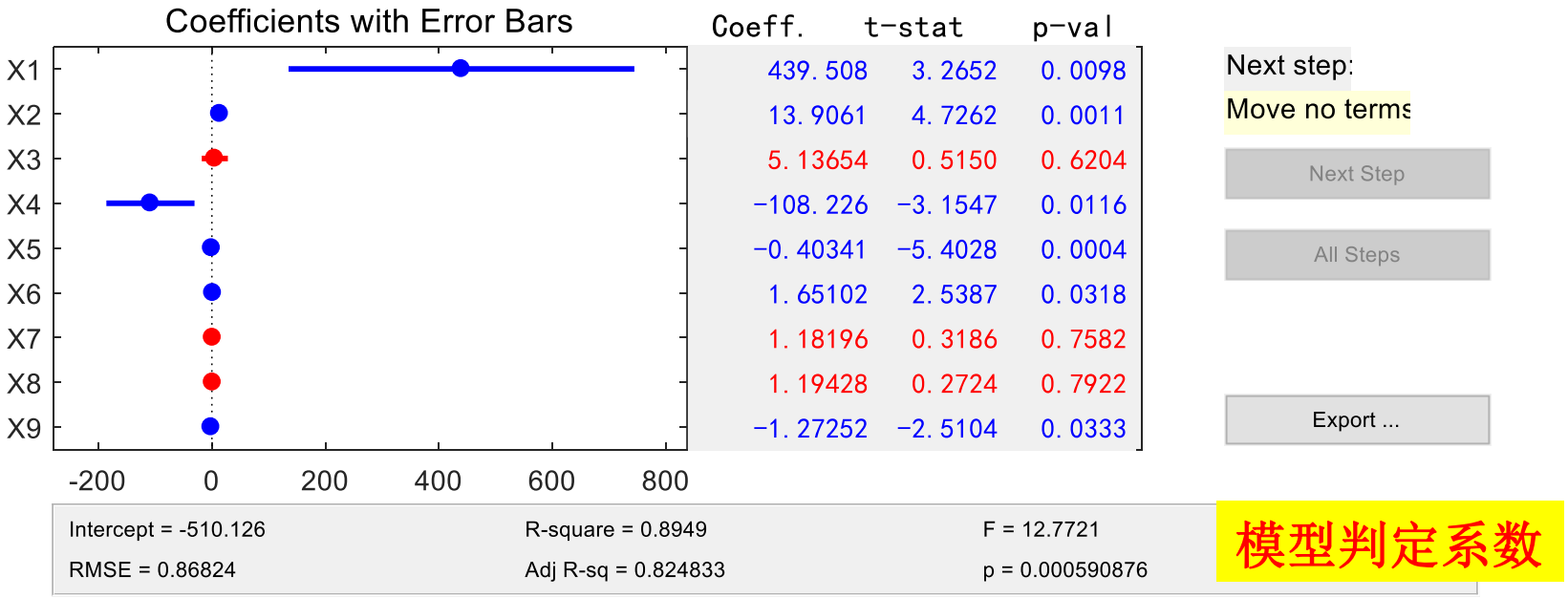
```
>> data = xlsread('light.xlsx');
>> x1 = [data(:,1);data(:,5)];
>> x2 = [data(:,2);data(:,6)];
>> x3 = [data(:,3);data(:,7)];
>> y = [data(:,4);data(:,8)];
>> model=[x1 x2 x3 x1.^2 x2.^2 x3.^2 x1.*x2
x1.*x3 x2.*x3];
>> stepwise(model,y,[1:9],0.05,0.05)
```

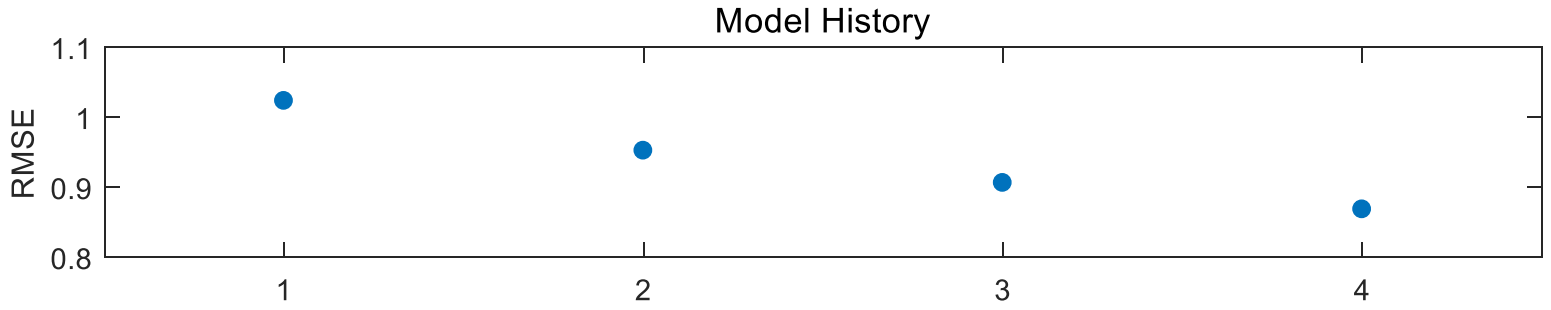
初始化界面

逐步回归函数移除变量

经过三步，最终结果界面

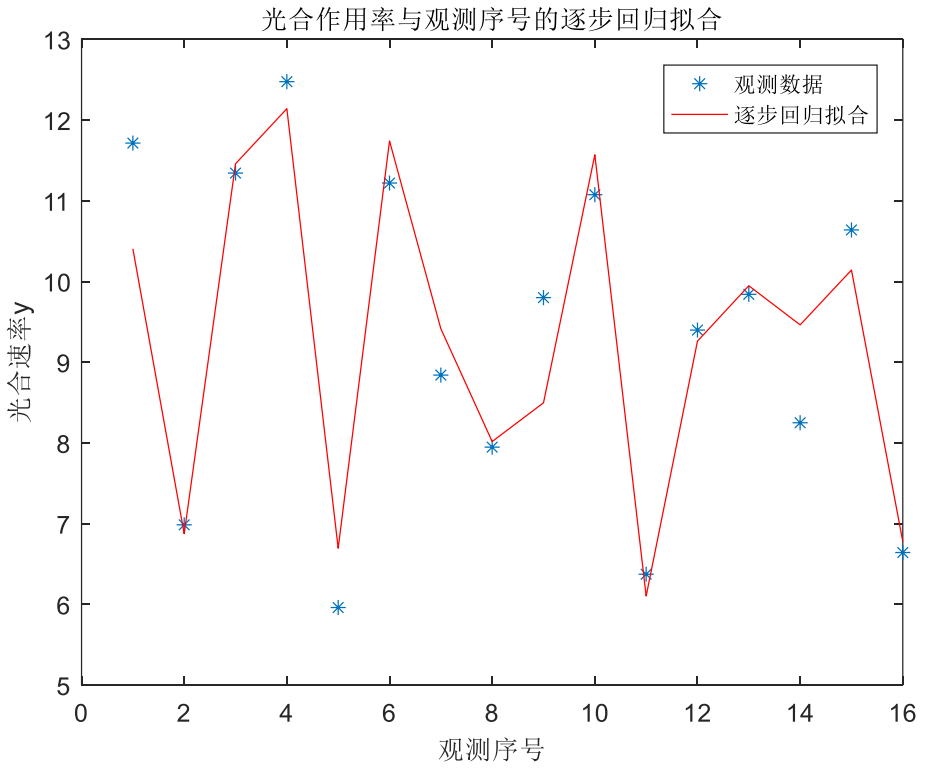
$$y = 439.508x_1 + 13.9061x_2 - 108.226x_1^2 - 0.40341x_2^2 + 1.65102x_3^2 - 1.27252x_2x_3 - 510.13$$





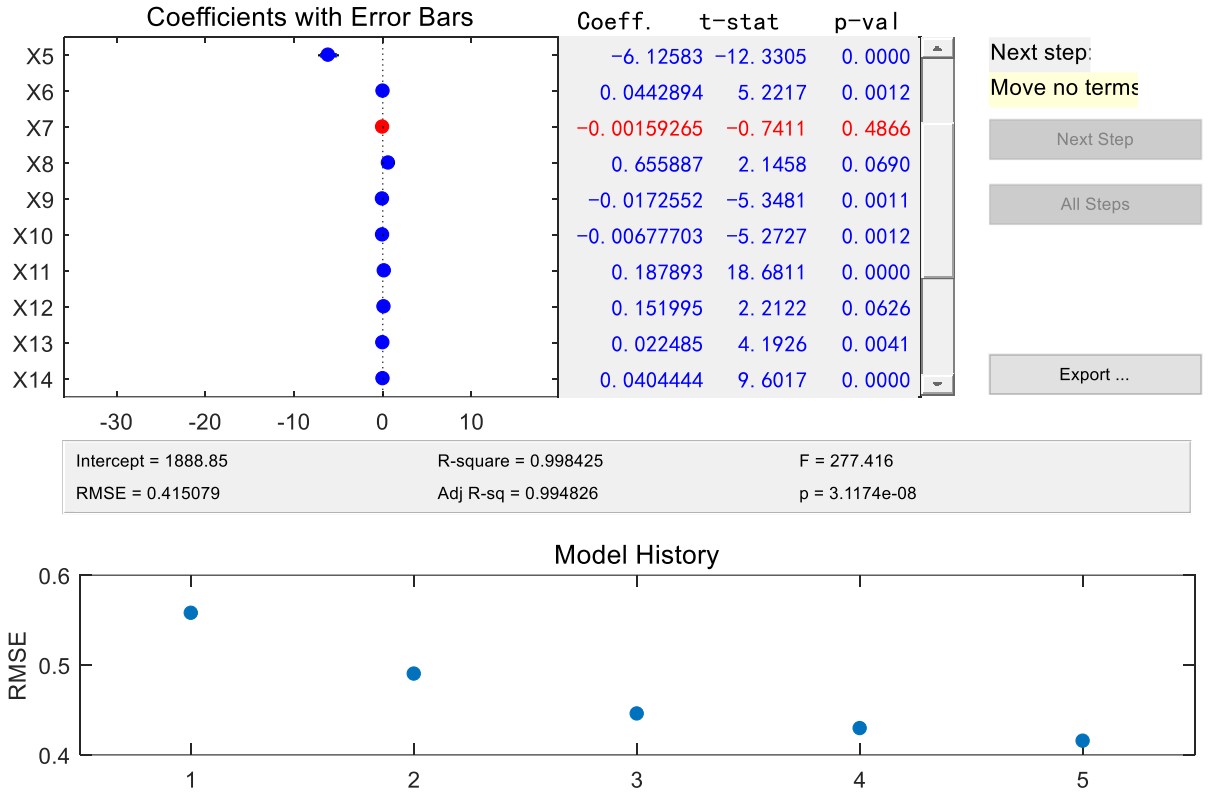
逐步回归函数回归曲线

```
yp= model *beta+stats.intercept;
plot(y,'*');
hold on;
plot(yp,'r')
xlabel('观测序号');
ylabel('光合速率y');
legend('观测数据','逐步回归拟合','Location','Northeast')
set(gca,'color','none')
```



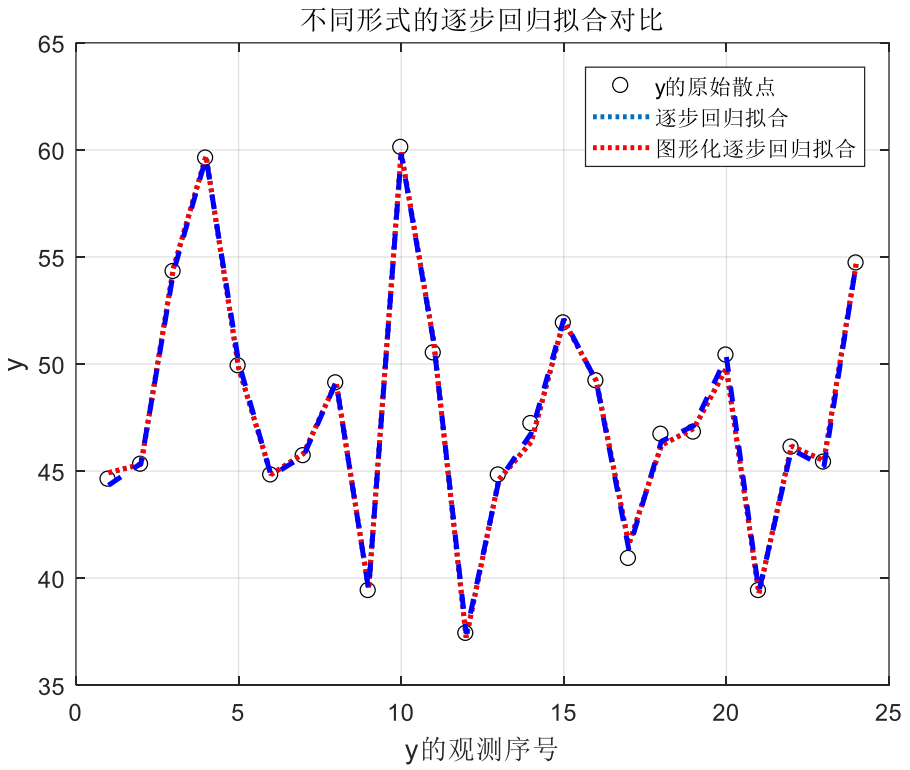
2. 逐步回归函数stepwiselm

```
>> data = xlsread('body.xls');
>> X = data(:,3:7);
>> y = data(:,2);
>> mds = LinearModel.stepwise(X,y,'poly22222')
>> lms = stepwiselm(X,y,'quadratic')    %另一种形式
>> x1 = X(:,1);
>> x2 = X(:,2);
>> x3 = X(:,3);
>> x4 = X(:,4);
>> x5 = X(:,5);
>> Xdata = [x1,x2,x3,x4,x5,x1.^2,x2.^2,x3.^2,x4.^2,x5.^2,x1.*x2,x1.*x3,x1.*x4,x1.*x5,x2.*x3,x2.*x4,x2.*x5,x3.*x4,x3.*x5,x4.*x5];
>> stepwise(Xdata,y,[1:20])
```



2. 逐步回归函数stepwiselm

```
>> yfitted = mds.Fitted;
>> plot(y,'ko');
>> hold on
>> plot(yfitted,'r:','linewidth',2);
>> xlabel('y的观测序号');
>> ylabel('y');
>> yp = Xdata*beta+stats.intercept;
>> plot(yp,'b--','linewidth',2);
>> grid on
>> legend('y的原始散点','逐步回归拟合','图形化逐步回归拟合');
>> title('不同形式的逐步回归拟合对比')
```



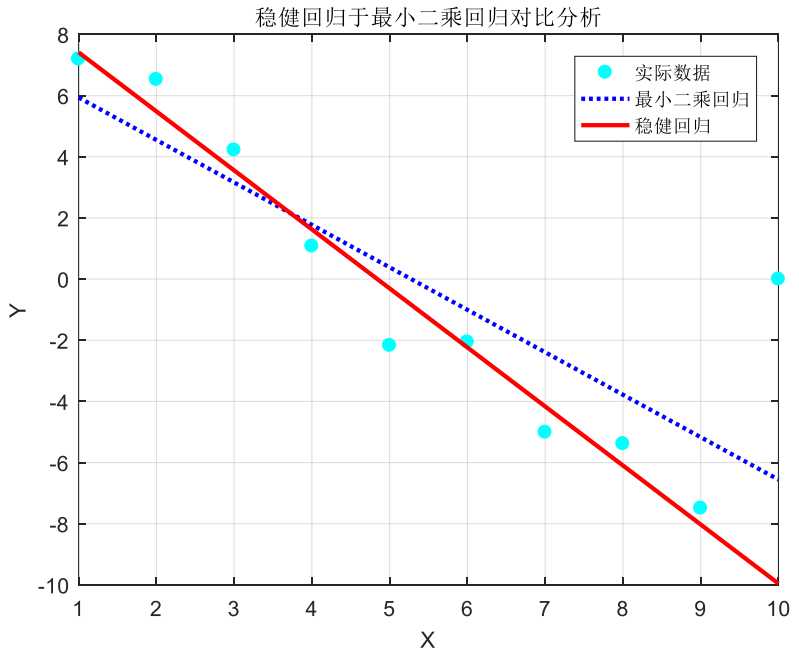
二. 稳健线性回归函数robustfit

- 稳健回归(Robust Regression)估计，如误差为正态时，它比最小二乘估计LSE稍差一点，但误差非正态时，它比LSE要好得多。这种对误差项分布的稳健特性，常能有效排除异常值干扰。
- 一般回归模型:

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i, i = 1, \cdots, n$$

其中β_j为未知回归系数，e_i独立同分布，均值为0。

最小二乘法是找β_j使表达式 $\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$ 达到最小作为代价函数。这样做会往往使得那些远离数据群体的数据（很可能是异常值）对残差平方和影响比其他数据大得多。这是因为最小二乘估计为了达到极小化残差平方和的目的，必须迁就远端的数据，所以异常值对于参数估计相当敏感。



二. 稳健线性回归函数robustfit

M估计稳健回归的基本思想是采用迭代加权最小二乘估计回归系数，根据回归残差的大小确定各点的权wi，以达到稳健的目的，其优化的目标函数是：

$$\sum_{i=1}^{n} w_i \left(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 = \min$$

为减少“异常点”作用，可以对不同的点施加不同的权重，即对残差小的点给予较大的权重，而对残差较大的点给予较小的权重，根据残差大小确定权重，并据此建立加权的最小二乘估计，反复迭代以改进权重系数，直至权重系数之改变小于一定的允许误差(tolerance)。其参数β_j可采用迭代加权最小二乘方法求解。

稳健回归对数据的拟合程度好些，忽略了异常值。最小二乘拟合则受到异常值的影响，向异常值偏移。

二. 稳健线性回归函数robustfit

- [b, stats] = robustfit(X,y,wfun,tune, const)
  - 输入参数 X 为 n×p 的自变量矩阵(或称预测变量矩阵, 设计矩阵), 对应 p个预测因子对 n 个观测值中每个的贡献。y 是 n×1 观测值向量(或称响应向量)，输出的 b 为 (p+1)×1 向量;
  - 缺省情况下，算法使用基于bisquare加权函数的迭代重加权最小二乘法。
  - 该函数增加了一个加权函数“wfun”和常数“tune”。“tune”是一个调节常数，其在计算权重之前被分成残差向量，如果“wfun”被指定为一个函数，那么“tune”是必不可少的。权重函数“wfun”可以为下表中的任何一个权重函数;
  - 参数const来控制模型中是否包含常数项，const取值为'on'(默认值)或'off'。

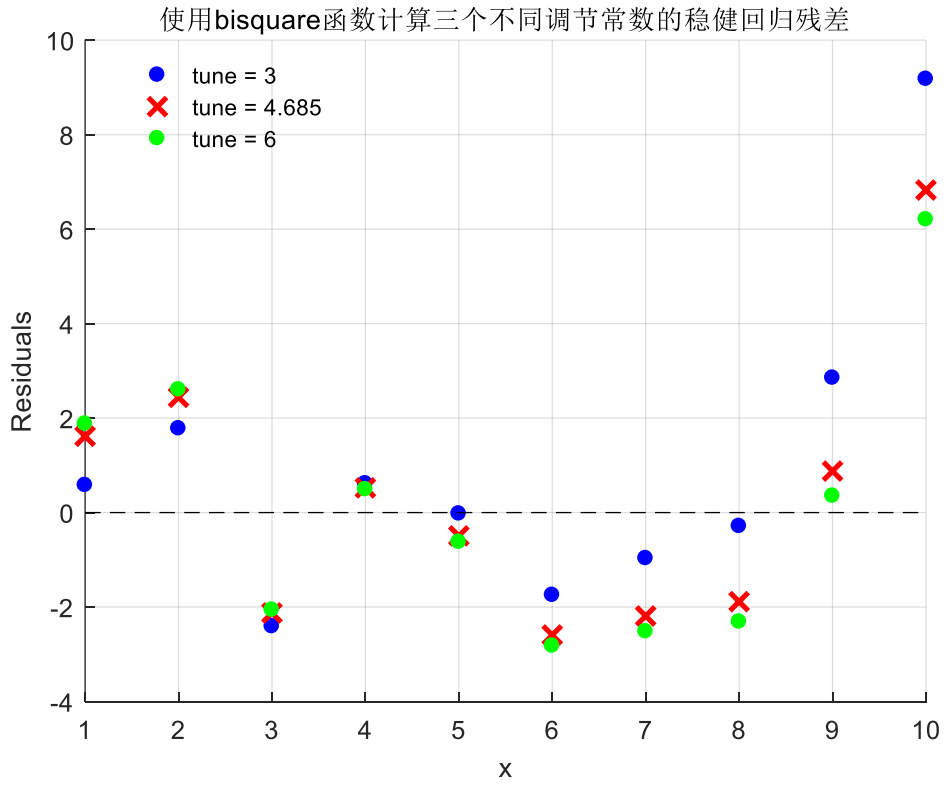
二. 稳健线性回归函数robustfit

Weight Function	Description	Default Tuning Constant
'andrews'	w = (abs(r)<pi) .* sin(r) ./ r $w = \sin(r)/r \ if \ \|r\| < \pi$	1.339
'bisquare'	w = (abs(r)<1) .* (1 - r.^2).^2 (also called biweight)	4.685
'cauchy'	w = 1 ./ (1 + r.^2)	2.385
'fair'	w = 1 ./ (1 + abs(r))	1.400
'huber'	w = 1 ./ max(1, abs(r))	1.345
'logistic'	w = tanh(r) ./ r	1.205
'ols'	Ordinary least squares (no weighting function)	None
'talwar'	w = 1 * (abs(r)<1)	2.795
'welsch'	w = exp(-(r.^2))	2.985
function handle	Custom weight function that accepts a vector r of scaled residuals, and returns a vector of weights the same size as r	1

r = resid/(tune*s*sqrt(1-h))，其中resid是上一次迭代的残差向量。tune是调谐常数。h是最小平方拟合的杠杆值向量。s 是由s=MAD/0.6745给出的误差项标准差的估计值。其中 MAD 为残差绝对值的中位数。常数0.6745保证了在正态分布下估计是无偏的。如果 X 中有 p 列，则在计算 MAD 时会将残差绝对值向量的前 p 个最小值舍去。

二. 稳健线性回归函数robustfit

```
x = (1:10)';
rng ('default') % For reproducibility
y = 10 - 2*x + randn(10,1); %加入随机干扰项
y(10) = 0;  %第十个数值修改为0，成为异常值
%使用bisquare函数计算三个不同调节常数的稳健回归残差。默认调节常数为4.685。
tune_const = [3 4.685 6];
for i = 1:length(tune_const)
    [~,stats] = robustfit(x,y,'bisquare',tune_const(i));
    resids(:,i) = stats.resid;
end
scatter(x,resids(:,1),'b','filled') %绘制残差
hold on
plot(resids(:,2),'rx','MarkerSize',10,'LineWidth',2)
scatter(x,resids(:,3),'g','filled')
plot([min(x) max(x)],[0 0],'--k') %绘制参考线
hold off; grid on
xlabel('x'); ylabel('Residuals')
legend('tune = 3','tune = 4.685','tune = 6','Location','best')
legend('boxoff')
```



二. 稳健线性回归函数robustfit

>> rmse = sqrt(mean(resids.^2)) %计算三个不同tune的均方根残差

rmse =

3.2577 2.7576 2.7099

%因为增加调整常数会降低分配给异常值的权重，所以RMSE会随着调整常数的增加而减小。

>> bls = regress(y,[ones(10,1) x]) %普通LS线性回归

bls =

7.8518 -1.3644

>> [brob,stats] = robustfit(x,y); %稳健回归

brob =

8.4504 -1.5278

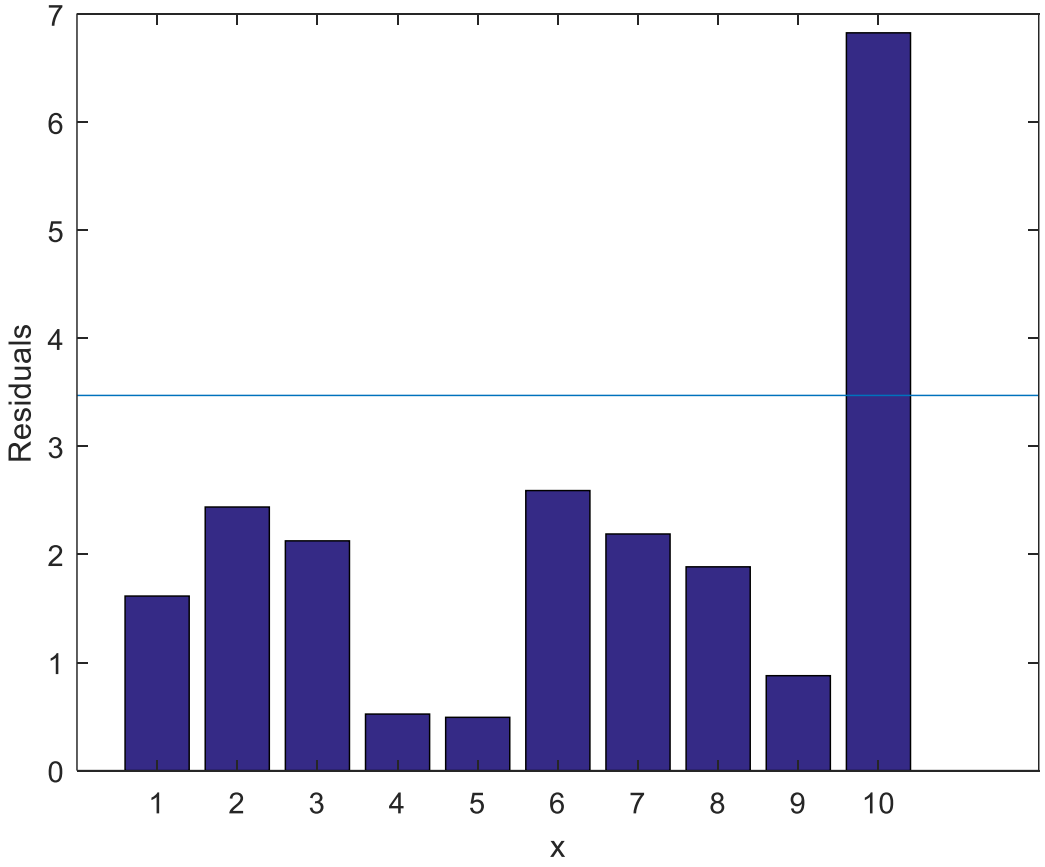
>> outliers_ind = find(abs(stats.resid)>stats.mad_s)

outliers_ind = 10 %异常点的行索引

bar(abs(stats.resid)) %绘制残差条形图

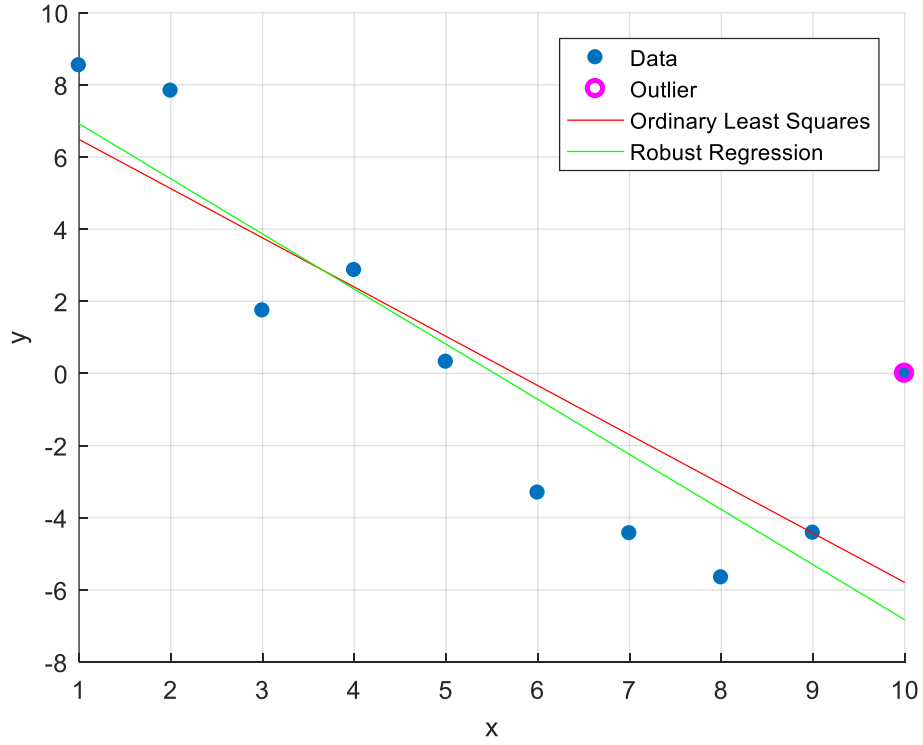
refline(0,stats.mad_s) %残差绝对值的中位数

xlabel('x');ylabel('Residuals')



二. 稳健线性回归函数robustfit

```
scatter(x,y,'filled')
hold on
plot(x(outliers_ind),y(outliers_ind),'mo','LineWidth',2) %异常点
plot(x,bls(1)+bls(2)*x,'r') %最小二乘直线
plot(x,brob(1)+brob(2)*x,'g')  %稳健回归直线
hold off
xlabel('x'); ylabel('y')
legend('Data','Outlier','Ordinary Least Squares','Robust Regression')
grid on
```



异常值对稳健拟合的影响小于最小二乘拟合。

三. 响应面分析函数rstool

- 许多工业试验中考察的指标（称为响应变量或因变量）经常受很多因素（称为因子变量或自变量）的影响。试验的目的是找出当这些因素取何值时，考察的指标最佳。
- 假定指标和因素间满足二次函数关系，如果每个因素测定三个以上不同值，那么二次曲面可以由最小二乘估计法得到；
  - 如果得到的曲面是凸面（像山丘）或凹面（像山谷）这类简单曲面，那么预测的最佳指标（极大值或极小值）可以从所估计的曲面上获得；
  - 如果曲面很复杂，或者预测的最佳点远离所考察因素的试验范围，那么可以通过分析二次曲面的形状，来确定重新进行试验的方向。

三. 响应面分析函数rstool

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

- 在响应面分析中，首先要得到回归方程：$\hat{y} = f(x_1, x_2, \cdots, x_l)$，然后通过对自变量$x_1, x_2, \cdots, x_l$的合理取值，求得使$\hat{y}$最优的值，这就是响应面分析的目的。
- 假定某个响应变量y在两个因子变量x_1和x_2的一些组合值上被测量，关于响应变量y的二次响应曲面回归模型为：

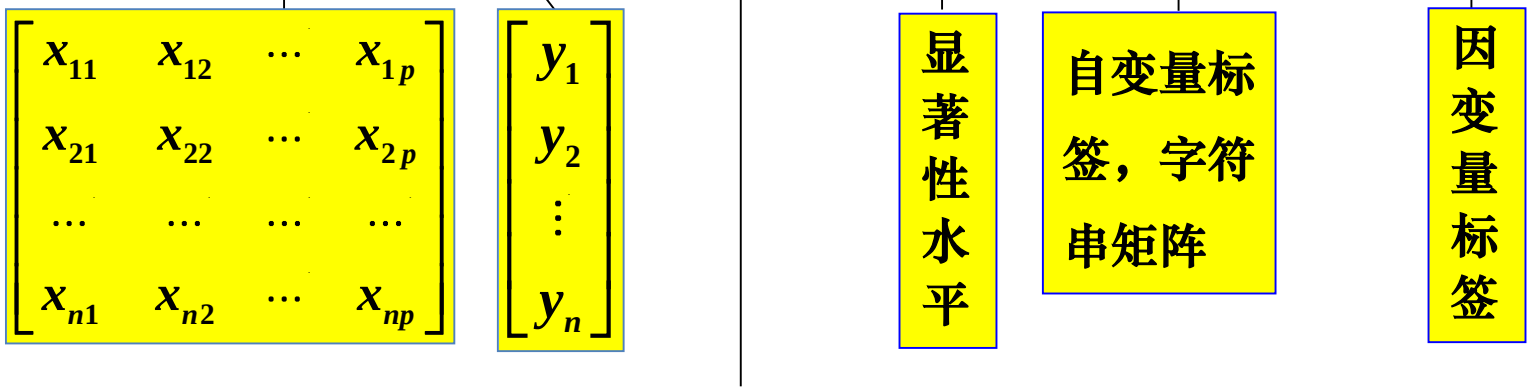
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$$

- 对这样的数据进行分析一般有以下三项任务：
  - 模型拟合及对参数估计作方差分析；
  - 为了调查预测响应曲面的形状而进行典型相关分析；
  - 为了寻找最佳响应的范围而进行岭嵴分析。

三. 响应面分析函数rstool

信阳师范学院
数学与统计学院

$$rstool(x,\quad y,\quad model, alpha, 'xname', 'yname')$$



由下列4个模型中选择1个（用字符串输入，缺省时为线性模型）：

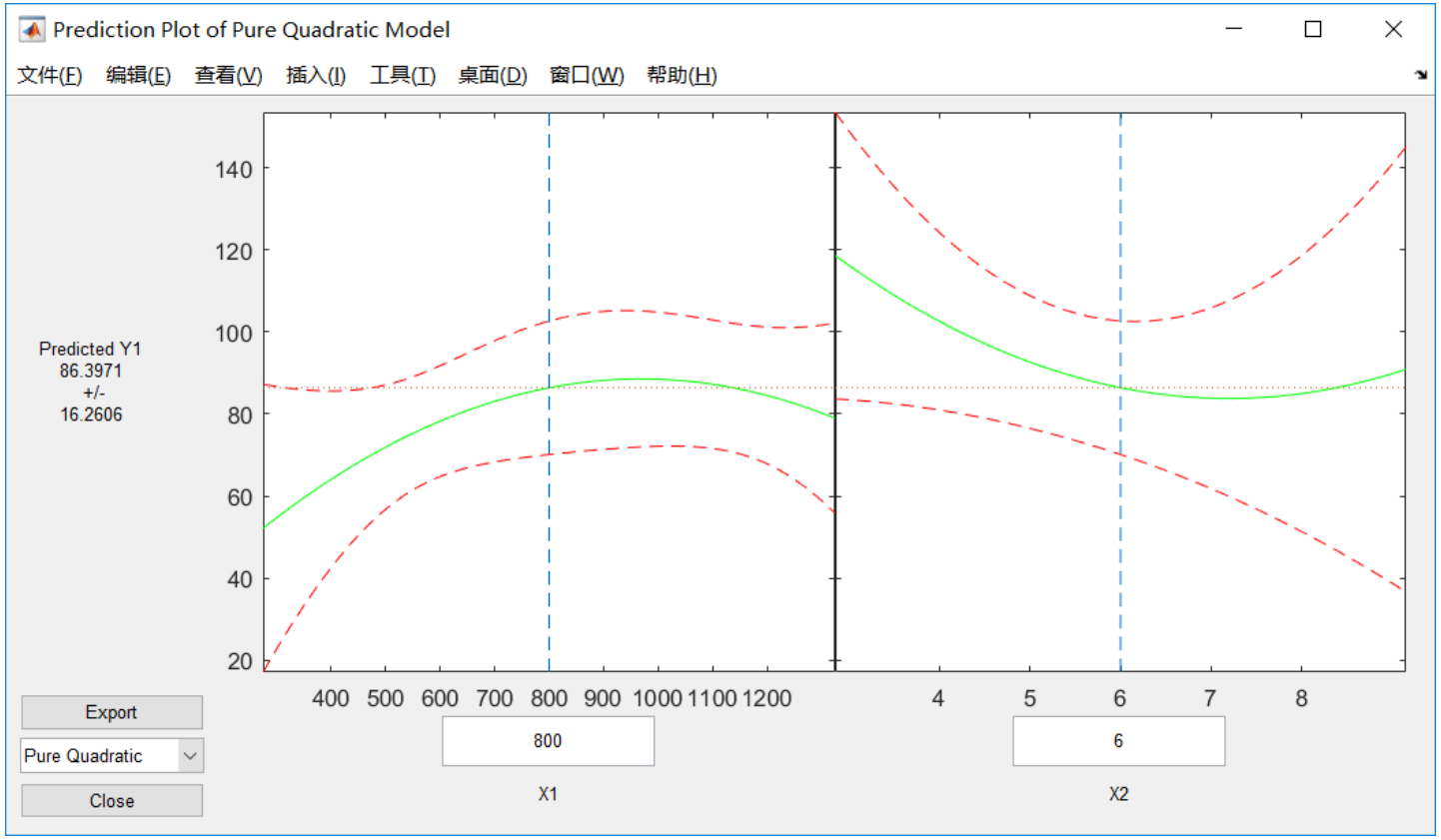
linear（线性）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$

purequadratic（纯二次）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^{n} b_{jj} x_j^2$

interaction（交叉）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \le j \ne k \le m} b_{jk} x_j x_k$

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

三. 响应面分析函数rstool



- rstool 的输出是一个交互式画面，画面中有m个图形，分别给出了一个独立变量x_i与y的拟合曲线，以及y的置信区间，此时其余$m-1$个变量取固定值。可以输入不同的变量的不同值得到y的相应值。
- 图的左下方有两个下拉式菜单，一个用于传送回归系数、剩余标准差、残差等数据；另一个用于选择四种回归模型中的一种，选择不同的回归模型，其中剩余标准差最接近于零的模型回归效果最好。

案例分析：需求量与收入价格

例：设某商品的需求量与消费者的平均收入、商品价格的统计数据如下，建立回归模型，预测平均收入为 1000、价格为 6 时的商品需求量。

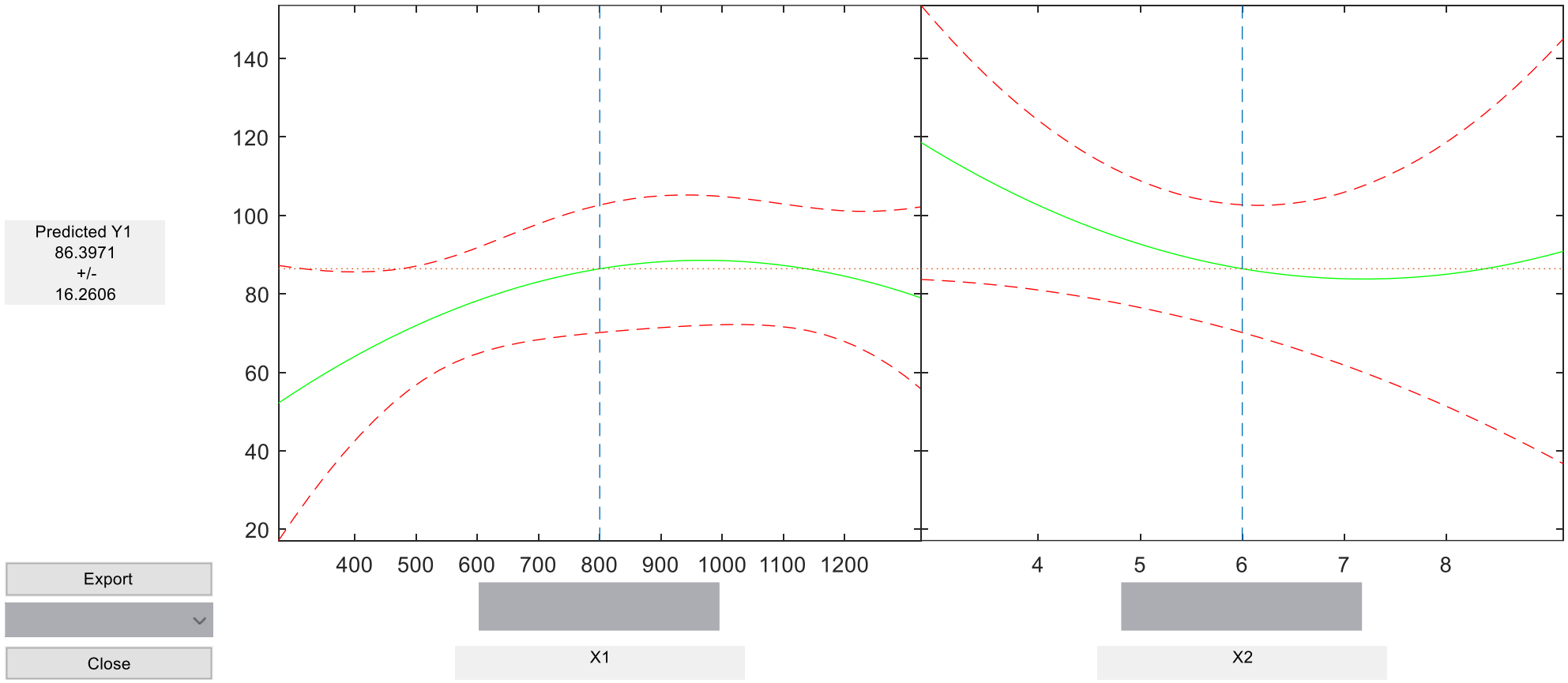
需求量	100	75	80	70	50	65	90	100	110	60
收　入	1000	600	1200	500	300	400	1300	1100	1300	300
价　格	5	7	6	6	8	7	5	4	3	9

选择纯二次模型，即 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$

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

案例分析：需求量与收入价格

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$



四. 曲线拟合工具箱函数cftool

- 曲线拟合应用程序提供了一个灵活的界面，可以在其中交互地将曲线和曲面拟合到数据和视图绘图。如
  - 可以创建、绘制和比较多个拟合；
  - 使用线性或非线性回归、插值、平滑和自定义公式；查看拟合优度统计数据，
  - 显示置信区间和残差，删除异常值，并根据验证数据评估拟合度；
  - 自动生成代码以拟合和绘制曲线和曲面，或将拟合导出到工作区以进行进一步分析。
- cftool(x, y [], W)： creates a curve fit to x input and y output. x and y must be numeric, have two or more elements, and have the same number of elements. cftool opens Curve Fitting app if necessary.
- cftool(X, Y, Z, W)： creates a surface fit with weights W. W must be numeric and have the same number of elements as Z.

案例分析：钢的强度和硬度

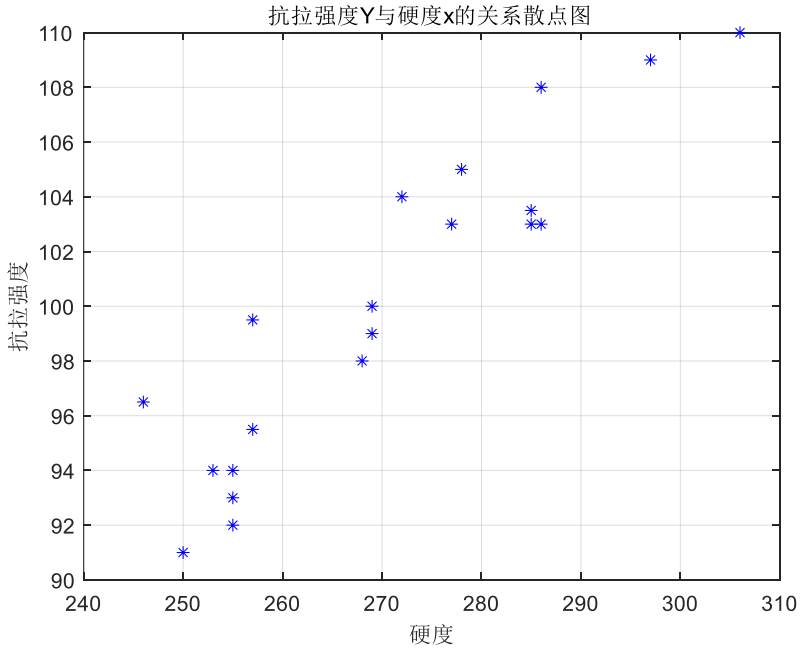
信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

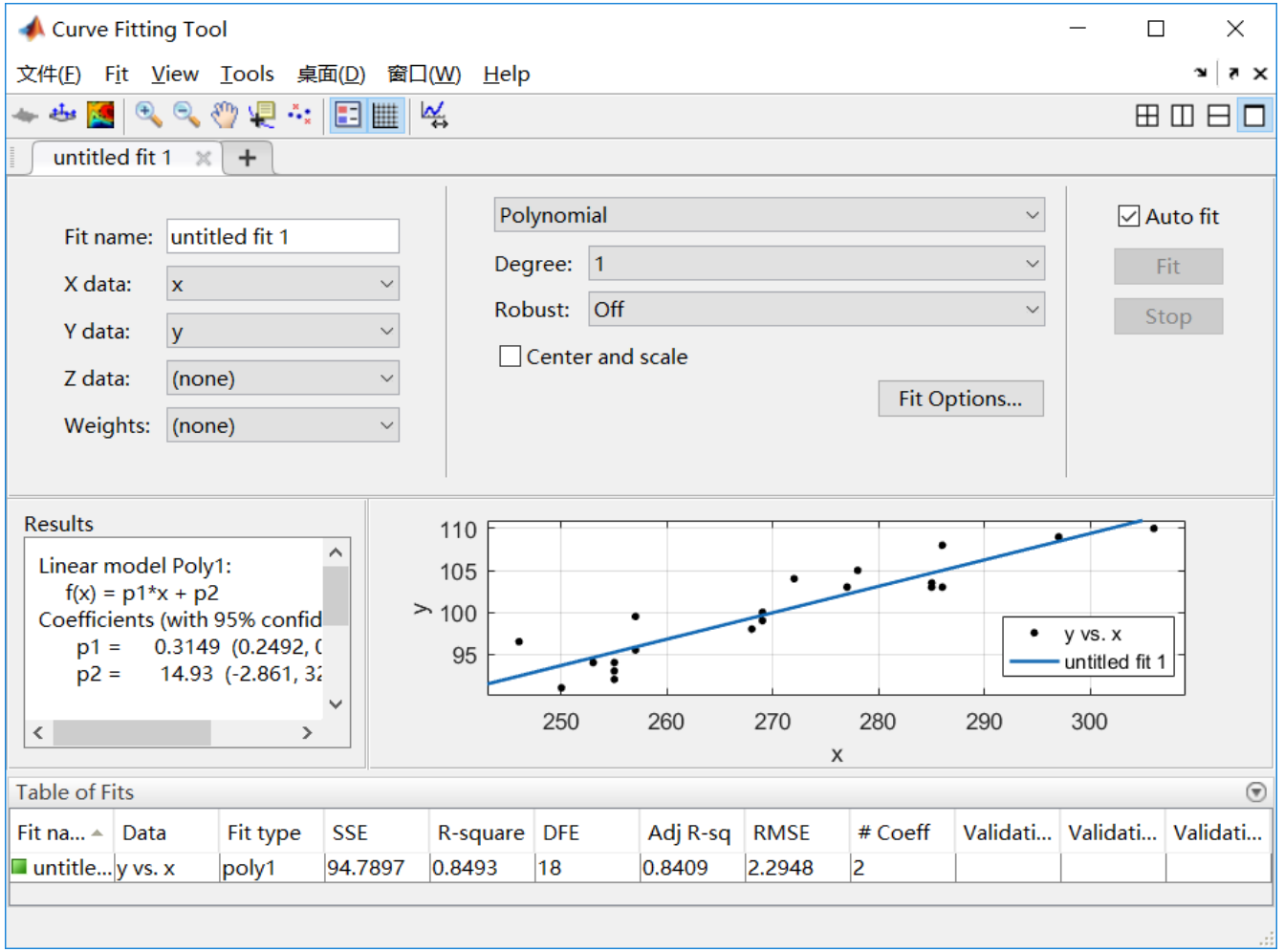
例：钢的强度和硬度都是反映钢质量的指标。现在炼20炉中碳钢，它们的抗拉强度Y与硬度x的20对实验值如下表。试绘出散点图，求Y对x的经验回归直线方程。

编号	x_i	y_i	编号	x_i	y_i	编号	x_i	y_i	编号	x_i	y_i
1	277	103	6	268	98	11	286	108	16	255	94
2	257	99.5	7	285	103.5	12	269	100	17	269	99
3	255	93	8	286	103	13	246	96.5	18	297	109
4	278	105	9	272	104	14	255	92	19	257	95.5
5	306	110	10	285	103	15	253	94	20	250	91

案例分析：钢的强度和硬度

```
data=[277,103;257,99.5;255,93;278,105;306,110;268,98;285,103.5;286,103;272,104;285,103;286,108;269,100;246,96.5;
255,92;253,94;255,94;269,99;297,109;257,95.5;250,91];
x= data(:,1);y= data(:,2);
plot(x,y,'*')
xlabel('硬度'); ylabel('抗拉强度');
set(gca,'color','none')
cftool(x,y)
```






01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听