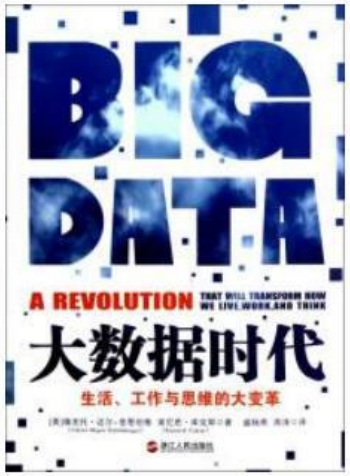
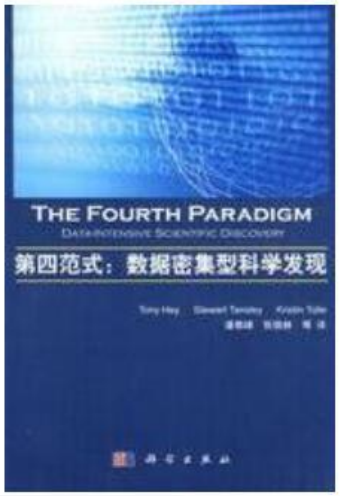
# 插值与拟合

钟秋平

广东岭南职业技术学院

# 大数据




身处
大数据时代

我们已经处在大数据时代

**时间就是生命！**

**事件一：变革公共卫生**

| 事件 | 2009年，H1N1流感预测 | |
|---|---|---|
| 对手 | 谷歌 | 疾控中心 |
| 武器 | 分析搜索记录 | 医院报告 |
| 结果 | 谷哥提前两周得到结果<br>与官方数据相关性达97% | |

**省钱是硬道理！**

**事件二：变革商业**

| 事件 | 机票价格预测 |
|---|---|
| 人物 | 埃齐奥尼的Farecast系统 |
| 武器 | 分析大量价格记录 |
| 结果 | 票价预测准确度达75%<br>平均每张机票节省50美元 |

# 数据分析

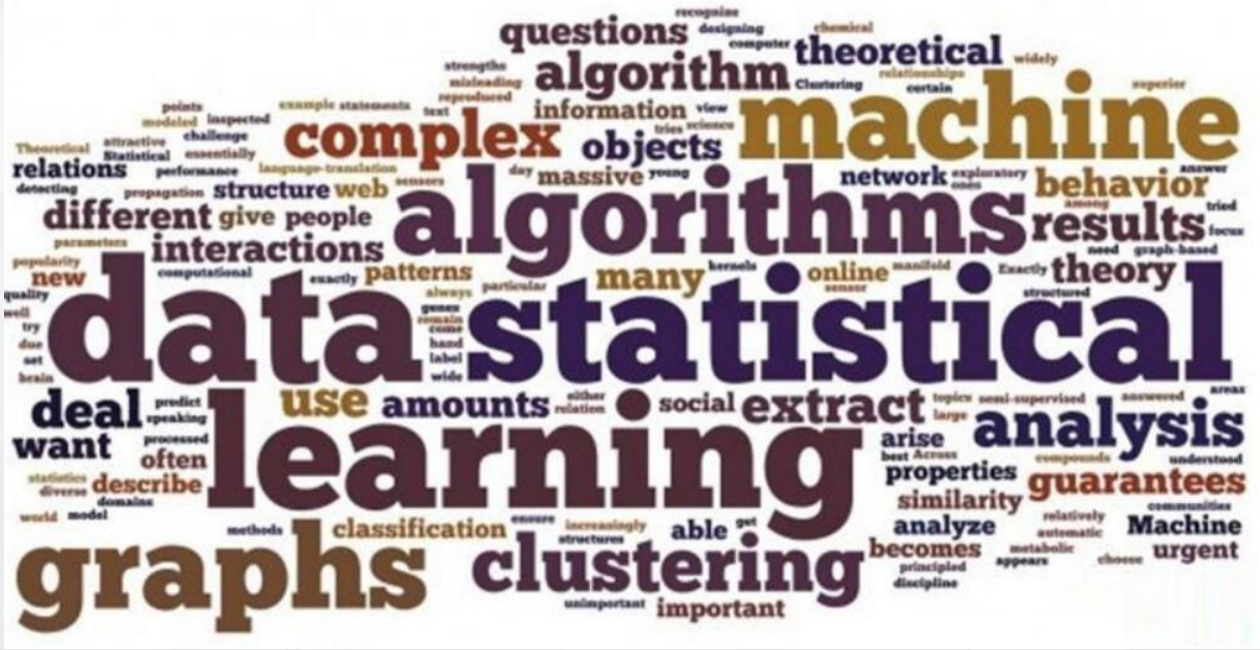
- **探索性数据分析**：当数据刚取得时，可能杂乱无章，看不出规律，通过作图、造表、用各种形式的方程拟合，计算某些特征量等手段探索规律性的可能形式，即往什么方向和用何种方式去寻找和揭示隐含在数据中的规律性。

- **模型分析**：在探索性分析的基础上提出一类或几类可能的模型，然后通过进一步的分析从中挑选一定的模型。

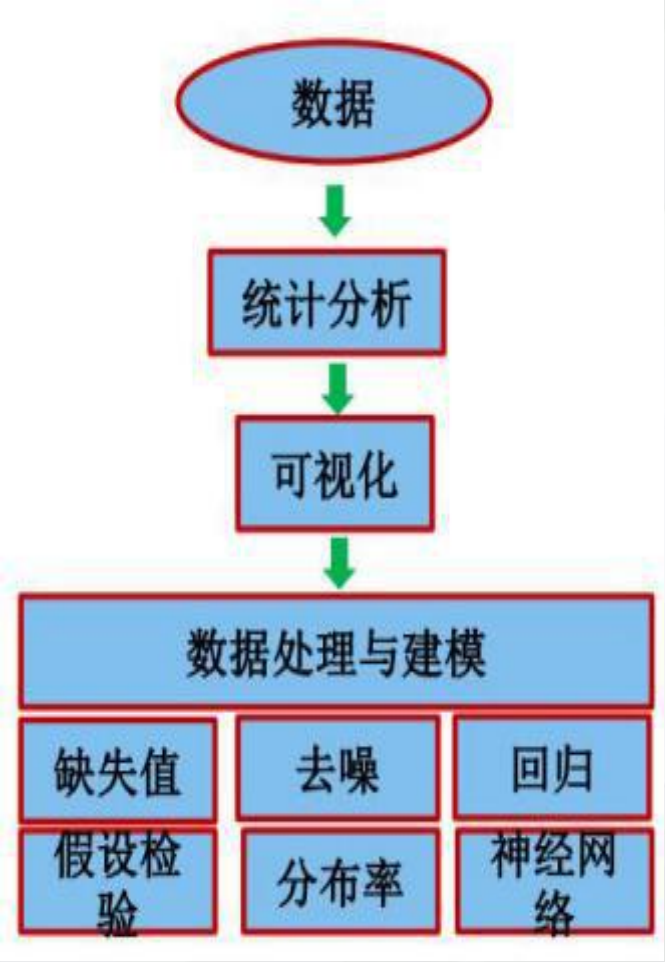
- **推断分析**：通常使用数理统计方法对所定模型或估计的可靠程度和精确程度作出推断。

# 数据模型常用的预测方法

- 插值、拟合方法
- 回归模型方法

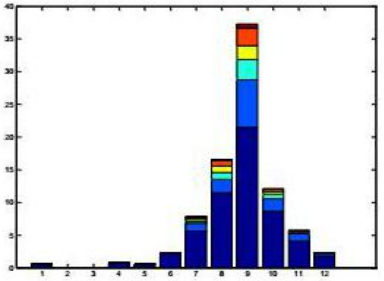


# 数据建模的流程




工作区

| 名称 | 值 | 大小 | 字节 | 类 | 最小值 | 最大值 | 极差 | 均值 |
|---|---|---|---|---|---|---|---|---|
| Australia | 19x1 double | 19x1 | 152 | double | NaN | NaN | NaN | NaN |
| Canada | 19x1 double | 19x1 | 152 | double | 1.3800 | 4.0800 | 2.7000 | 2.0868 |
| France | 19x1 double | 19x1 | 152 | double | 3.4100 | 7.5100 | 4.1000 | 4.4079 |
| Germany | 19x1 double | 19x1 | 152 | double | 2.6500 | 7.7500 | 5.1000 | 4.2247 |
| Italy | 19x1 double | 19x1 | 152 | double | 3.5700 | 7.6300 | 4.0600 | 4.6458 |
| Japan | 19x1 double | 19x1 | 152 | double | 2.8200 | 5.7400 | 2.9200 | 3.8205 |
| Mexico | 19x1 double | 19x1 | 152 | double | 1 | 2.4500 | 1.4500 | 1.7816 |
| SouthKorea | 19x1 double | 19x1 | 152 | double | 2.0500 | 6.2100 | 4.1600 | 3.8358 |
| UK | 19x1 double | 19x1 | 152 | double | 2.8200 | 7.4200 | 4.6000 | 4.3926 |
| USA | 19x1 double | 19x1 | 152 | double | 1.0600 | 3.2700 | 2.2100 | 1.5921 |
| Year | 19x1 double | 19x1 | 152 | double | 1990 | 2008 | 18 | 1999 |



$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 \sin(\beta_5 x + \beta_6)$$

# 应用案例

- 12名学生的英语、数学、物理、化学和政治成绩如表：

- 各科平均成绩
- 各科最高分及学生编号
- 各科最低分及学生编号

| 英语 | 数学 | 物理 | 化学 | 政治 |
|---|---|---|---|---|
| 90 | 93 | 95 | 96 | 77 |
| 90 | 67 | 85 | 74 | 67 |
| 89 | 99 | 72 | 89 | 59 |
| 78 | 88 | 90 | 73 | 71 |
| 56 | 66 | 58 | 79 | 83 |
| 67 | 78 | 82 | 80 | 90 |
| 88 | 78 | 67 | 78 | 86 |
| 65 | 75 | 79 | 81 | 74 |
| 73 | 63 | 85 | 76 | 67 |
| 76 | 86 | 91 | 89 | 57 |
| 90 | 97 | 73 | 68 | 86 |
| 71 | 83 | 87 | 78 | 77 |

# 案例分析

```
s=[90  93   91  85  77;
90   67   85  74    67;
89   99 72   89     59;
78 88     95 73     71;
56  66    58  79    83;
67  78    82  80    90;
88  78    67  78    86;
65  75    79   81    74;
73   63   85   76    67;
76   86   91   89    57;
91   97   73   68   86;
71  83    87   78   77]
```

```
s_sum=sum(s)   %求各科总成绩

s_avg=s_sum/12   %求各科平均成绩

s_max=max(s)   %求各科最高成绩

s_min=min(s)   %求各科最低成绩
```

```
[s_max,s_max_stu]=max(s)   %各科最高成绩,及对应的学生编号

[s_min,s_min_stu]=min(s)   %各科最高成绩,及对应的学生编号
```

# 案例分析

```
s=[90  93   91  85  77;
90   67   85  74    67;
89   99 72   89     59;
78 88     95 73     71;
56  66    58  79    83;
67  78    82  80    90;
88  78    67  78    86;
65  75    79   81   74;
73   63   85   76   67;
76   86   91   89   57;
91   97   73   68   86;
71  83    87   78   77]
```

```
s_std=std(s)  %求各科平均成绩
```

```
s_var=var(s)   %求各科最高成绩
```

```
s_stu_max=max(s')   %各个学生 5 科成绩总分
```

# 数据的拟合与插值

在某些问题中，由实验或测量易得到大量的数据。数据处理的目的是寻找数据内在的关系、规律并对未知的情形作出预测与预报。

插值与拟合都是根据一组数据构造一个近拟函数。

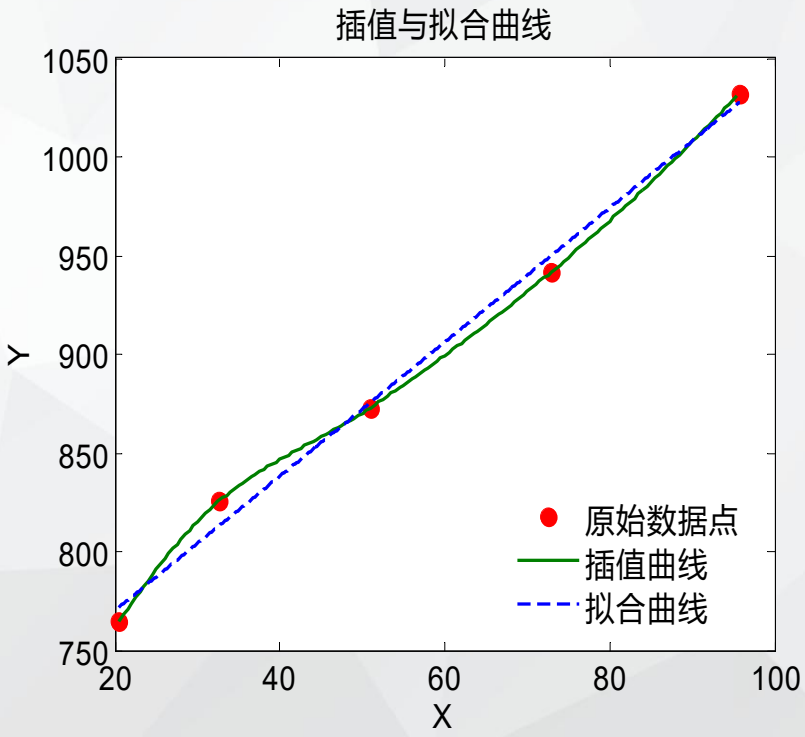
# 拟合与插值的定义

已知 $n+1$ 个不同点（$x_i$，$y_i$），$i$=0,1,2,...,$n$。

**插值**：构造一个通过全部点的函数 f($x$)（写不出具体的函数），然后求任意点 $x$ 的函数值 $y$。

**拟合**：构造一个函数 $f(x)$，使之在某种准则下与所有数据点最为接近，常用最小二乘法。

不一定经过所有的点



# 数据插值

**Remark:** 一元插值函数interp1的基本调用格式为

interp1(*x*, *y*, *cx*, 'method')

其中 *x*，*y* 分别表示已知数据点的横、纵坐标，*cx*为插值的横坐标，method为可选参数，可为

1）nearest——最近邻点插值

2）linear——线性插值（可缺省）

3）spline——三次样条插值

4）cubic——三次插值

# 插值模型——温度的推测

**问题提出**：在一天24小时内，从零点开始每间隔2h测得的温度数据如表1所示：

表1 温度测量数据 单位（°C）

| $t$ | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 12 | 9 | 9 | 10 | 18 | 24 | 28 | 27 | 25 | 20 | 18 | 15 | 13 |

**推测中午1点（即13点）是的温室？**

# 模型分析

画散点图，见图1

从图中观察测量数据规律不明显。很难找到函数，所以用插值法。

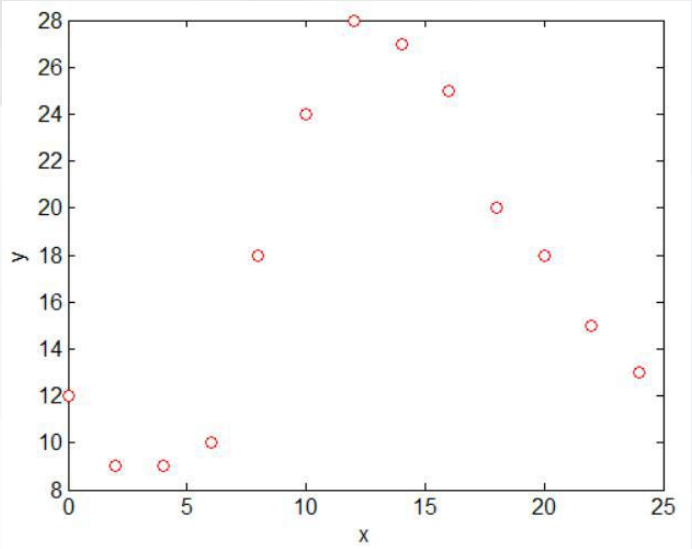



图1 温度测量数据散点图

Matlab代码

```
x=0:2:24;
y=[12 9 9 10 18 24 28 27 25 20 18 15 13];
plot(x,y,'.')
```

# 模型求解

用插值函数interp1( )推测中午13点的温度.

1)nearest插值

cy=27

Matlab代码

```
x=0:2:24;
y=[12 9 9 10 18 24 28 27 25 20 18 15 13];
cx=13;
cy=interp1(x,y,cx,'nearest')
```

2) linear插值

cy=27.5

Matlab代码

```
x=0:2:24;
y=[12 9 9 10 18 24 28 27 25 20 18 15 13];
cx=13;
cy=interp1(x,y,cx,'linear')
```

# 模型求解

用插值函数interp1( )推测中午13点的温度.

3) spline插值

cy=27.8725

Matlab代码

```
x=0:2:24;
y=[12 9 9 10 18 24 28 27 25 20 18 15 13];
cx=13;
xy=interp1(x,y,cx,'spline')
```

4) cubic插值

cy=27.6667

Matlab代码

```
x=0:2:24;
y=[12 9 9 10 18 24 28 27 25 20 18 15 13];
cx=13;
cy=interp1(x,y,cx,'cubic')
```

# 练习

[温度预测模型] 在12小时内，每隔一小时测量一次温室温度，具体数据见表2

表2 温室测量温度 单位（°C）

| 小时t | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 温度y | 5 | 8 | 9 | 15 | 25 | 29 | 31 | 30 | 22 | 25 | 27 | 24 |

求：温室在时间3.2、6.5、7.1、11.7对应的温度值。

**4.2** 缺失数据的预处理：

第十三层的缺失数据：由于在第一次和第二次的观测数据中，第十三层缺少一个点的观测数据，使得在寻找第十三层中心点时产生较大误差。因此，我们结合十二层与十一层第 5 个观测点坐标的相对变化情况，对第十三层的缺失数据进行了合理地赋值。根据对古塔各观测点散点图观察可见，古塔相邻两层的对应观测点坐标之间具有类似的关系。通过计算可得第一次测量中第十二层第 5 个观测点相对于第十一层第 5 个点的坐标变化值为(–0.055,0.173,4.271)，从而由第十二层第 5 个观测点坐标加上相对变化值可将第十三层的缺失数据赋值为(567.984,519.588,52.984)。同理可将第二次测量中第十三层的缺失数据赋值为(567.99,519.5816,52.983)。

塔尖的数据：在后两次测量中，塔尖仅有一个观测数据。由于塔尖各点坐标变化很小，所以对于只有一个测量点的塔尖数据，我们将其近似处理为塔尖中心点坐标。

# 数据的拟合

数据拟合与插值不同在于找出近似函数。先画数据散点图，根据其分布的总趋势剔除观察数据中的偶然误差，即数据修匀（数据光滑）问题。

常用拟合曲线的类型

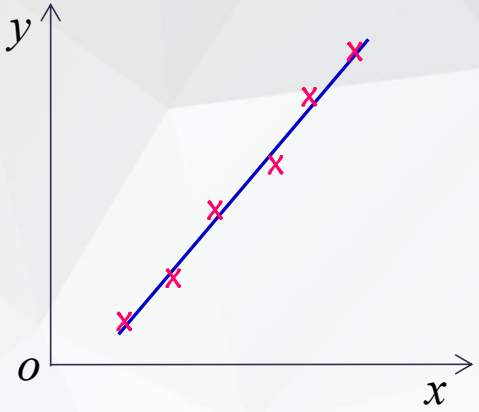
1）线性拟合：一次直线函数

2）曲线拟合：指数函数、对数函数、三角函数、二次函数、三次函数等高次多项式

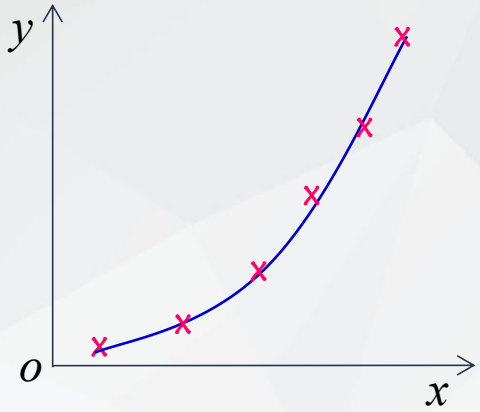
3）分段拟合：在不同范围内用不同函数

# 常用的拟合曲线

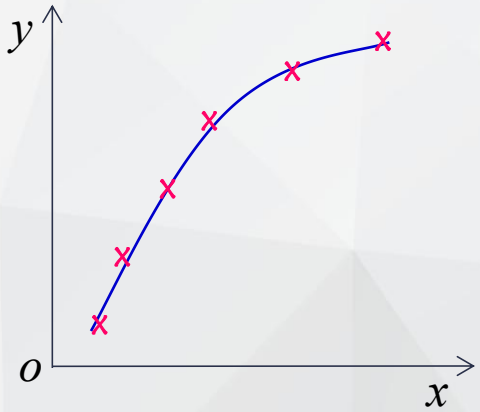
先画数据的散点图，通过观察选择几种合适的曲线分别拟合。



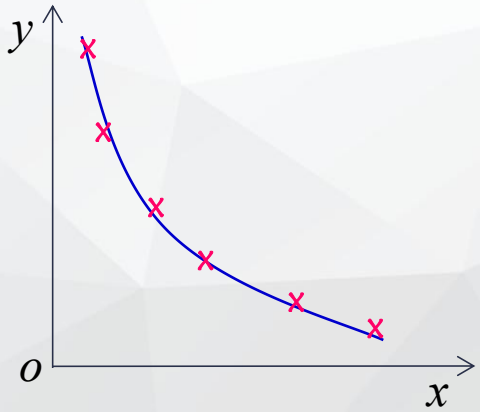

$y=ax+b$



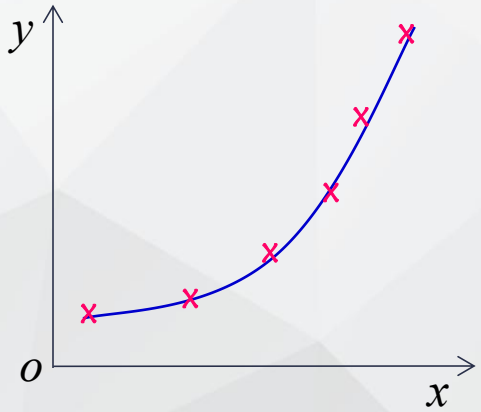

$y=ax^2+bx+c$




$y=ax^2+bx+c$




$y=a/x+b$




$y=ae^{bx}$

# 一元多项多拟合命令　polyfit

[p,s]=polyfit(*x*, *y*, *m*)

多项式是按次数从高到低排，如： $y=ax+b$, $y=ax^2+bx+c$

*m* 是多项式的最高次数，*x*, *y*为已知的数据。

输出的p表示系数，s用来估计预测误差.

求polyfit所得的拟合多项式在x处的预测值f

f=polyval(p,x)

# 模型——温度与电阻的关系

**问题提出**：有一个对温度敏感的电阻，现测得一组温度 $x$ 与电阻 $y$ 的数据见表3

表3 温度与电阻测量数据

| $x$ | 20.5 | 32.7 | 51.0 | 73 | 95.7 |
|---|---|---|---|---|---|
| $y$ | 765 | 826 | 873 | 942 | 1032 |

1）找出温度与电阻之间的函数关系

2）预测温度为60度时的电阻值

# 模型分析

画散点图，见图2

观察图不难发现，散点基本上在一条直线上，因此可用一次直线函数拟合。

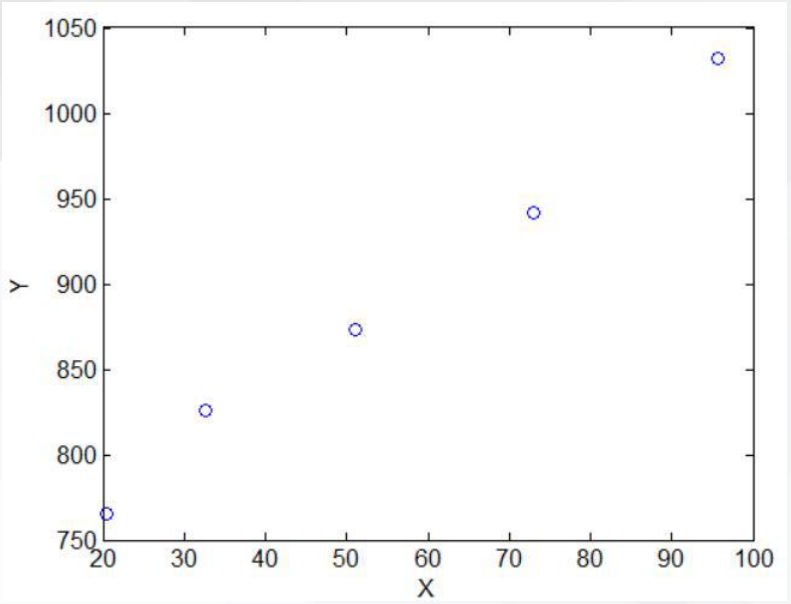



图2 温度与电阻散点图

设温度 $x$ 与电阻 $y$ 的拟合函数模型为 $y = ax + b$

# 模型建立与求解方法2 polyfit

```
%命令窗口输入
x=[20.5 32.7 51 73 95.7];
y=[765 826 873 942 1032];
[p,s]=polyfit(x,y,1)
```

```
%输出结果
p=
3.3987    702.0968
```

```
s =
R: [2x2 double]
df: 3
normr: 17.4235
```

$$\therefore y = 3.3987x + 702.0968$$

```
f=polyval(p,60)
```

```
f=polyval(p,x)
```

```
%x=60的预测值
f =906.0212
```

```
f =771.8    813.2    875.4    950.2    1027.4
%每个已知x点对应的预测值
```

# 模型检验

把拟合函数曲线图画在散点图上作比较，见图3。

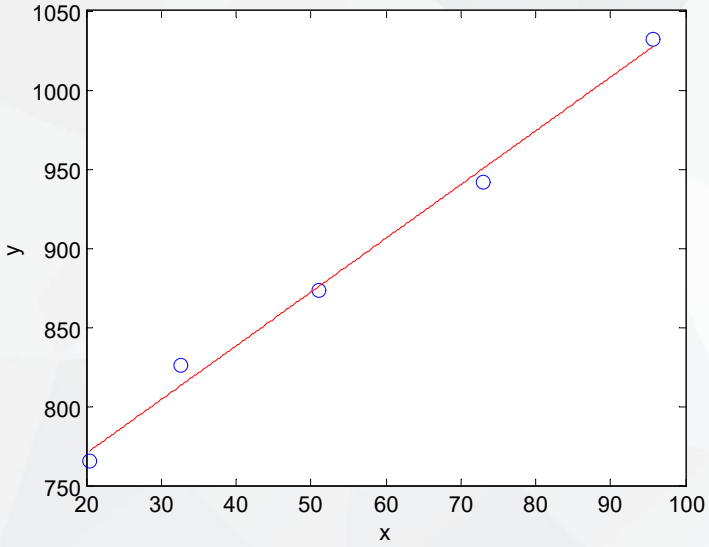



图3 温度与电阻拟合函数曲线图

```
x=[20.5 32.7 51 73 95.7];
y=[765 826 873 942 1032];
plot(x,y,'o')
hold on
x=20.5:0.1:95.7;
y=3.3987.*x+702.0968;
plot(x,y)
```

Matlab代码

# 模型2——消费水平与GDP的关系

问题提出：为研究人均消费水平，收集人均国内生产总值（人均GDP）$x$与人均消费金额$y$数据如表5所示。

表4　人均消费与人均GDP数据（单位：元）

| GDP | 4854 | 5576 | 6054 | 6308 | 6551 | 7086 | 7651 | 8214 | 9101 |
|---|---|---|---|---|---|---|---|---|---|
| 人均消费 | 2236 | 2641 | 2834 | 2972 | 3138 | 3397 | 3609 | 3818 | 4089 |

请根据表4数据，分析人均GDP与人均消费水平的关系。

# 模型分析

画散点图，见图4

观察图不难发现，可用一次函数、二次函数拟合。

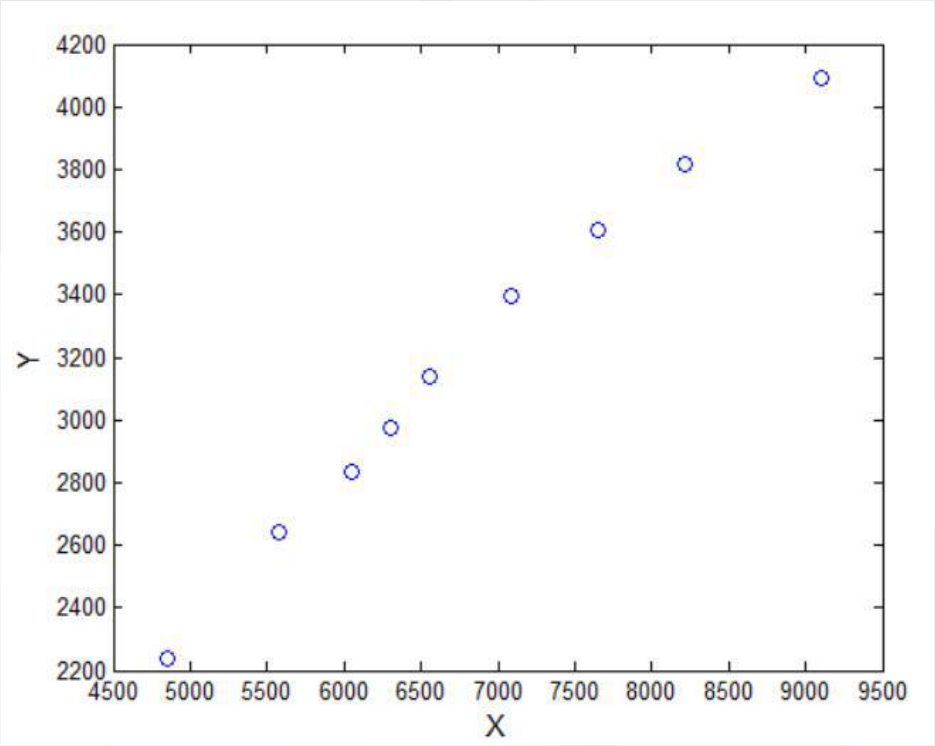



图4 人均GDP与人均消费散点图

# 模型建立与求解方法1———一次函数拟合

设人均DDP $x$ 与人均消费 $y$ 的线性模型为

$$y = ax + b$$

MATLAB命令polyfit( )求解得

$$y = 0.4414x + 181.6263$$

p=0.4414  181.6263

```
x=[4854 5576 6054 6308 6551 7086 7651 8214 9101];
y=[2236 2641 2834 2972 3138 3397 3609 3818 4089];
[p,s]=polyfit(x,y,1)
```

```
s =
R: [2x2 double]
df: 7
normr: 186.4815
```

# 模型建立与求解方法2——二次函数拟合

设人均DDP $x$ 与人均消费 $y$ 的二次函数模型为

$$y = ax^2 + bx + c$$

MATLAB命令polyfit( )求解得

$$y = -0.0000346x^2 + 0.926x - 1457.661$$

p= -0.0000346  0.926  -1457.661

```
format long
x=[4854 5576 6054 6308 6551 7086 7651 8214 9101];
y=[2236 2641 2834 2972 3138 3397 3609 3818 4089];
[p,s]=polyfit(x,y,2)
```

```
s =
R: [3x3 double]
df: 6
normr: 71.115
```

# 模型检验

二次函数拟合曲线图见图5。

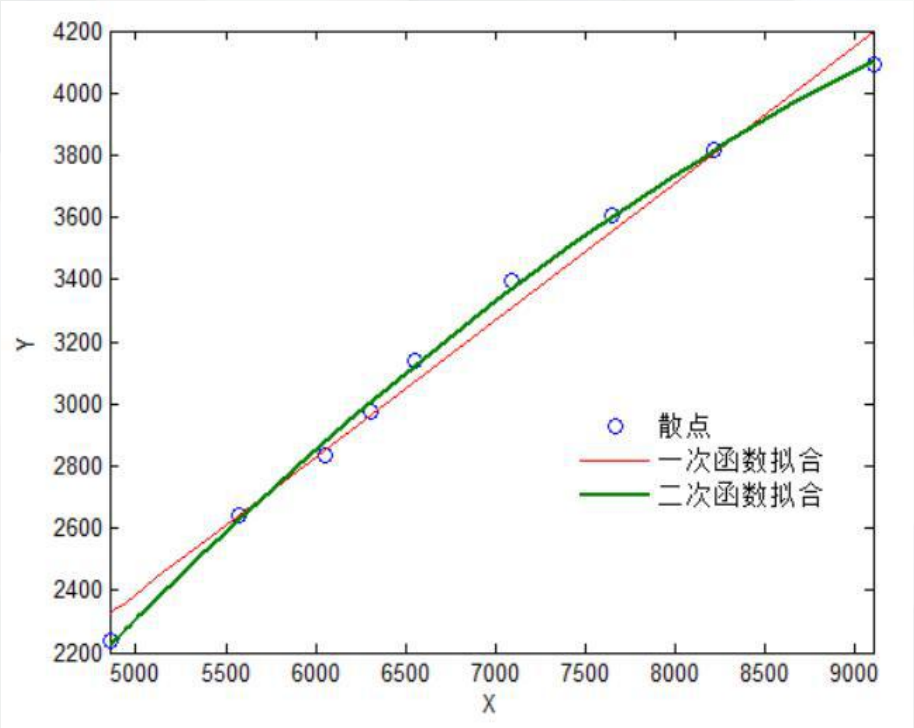



图5 人均GDP与人均消费拟合函数图

```
x=[4854 5576 6054 6308 6551
7086 7651 8214 9101];
y=[2236 2641 2834 2972 3138
3397 3609 3818 4089];
plot(x,y,'o')
hold on
fplot('0.4414*x+181.6263',[4854,
9101])
hold on
fplot('-0.0000346*x^2+0.926*x-
1457.661',[4854,9101])
```

# 模型检验

两种拟合曲线计算拟合值与实际值比较。

| x: GDP | 4854 | 5576 | 6054 | 6308 | 6551 | 7086 | 7651 | 8214 | 9101 |
|---|---|---|---|---|---|---|---|---|---|
| 人均消费y | 2236 | 2641 | 2834 | 2972 | 3138 | 3397 | 3609 | 3818 | 4089 |
| 一次拟合y1 | 2324.18 | 2642.87 | 2853.86 | 2965.98 | 3073.24 | 3309.39 | 3558.78 | 3807.29 | 4198.81 |
| 二次拟合y2 | 2221.92 | 2629.94 | 2880.22 | 3006.78 | 3123.69 | 3366.66 | 3601.76 | 3814.05 | 4104.01 |

$$J_2 = \sum_{i=1}^{n} [f(x_i) - y_i]^2 \quad = 5057.3 \ < J_1$$

二次函数拟合明显比一次拟合效果更好。

# 练习

【产量与施肥量的关系】在农业生产的试验研究中，对某地区土豆的产量与化肥的关系做了一系列实验，得到每公顷地氮肥的施肥量与土豆产量的关系如表4所示

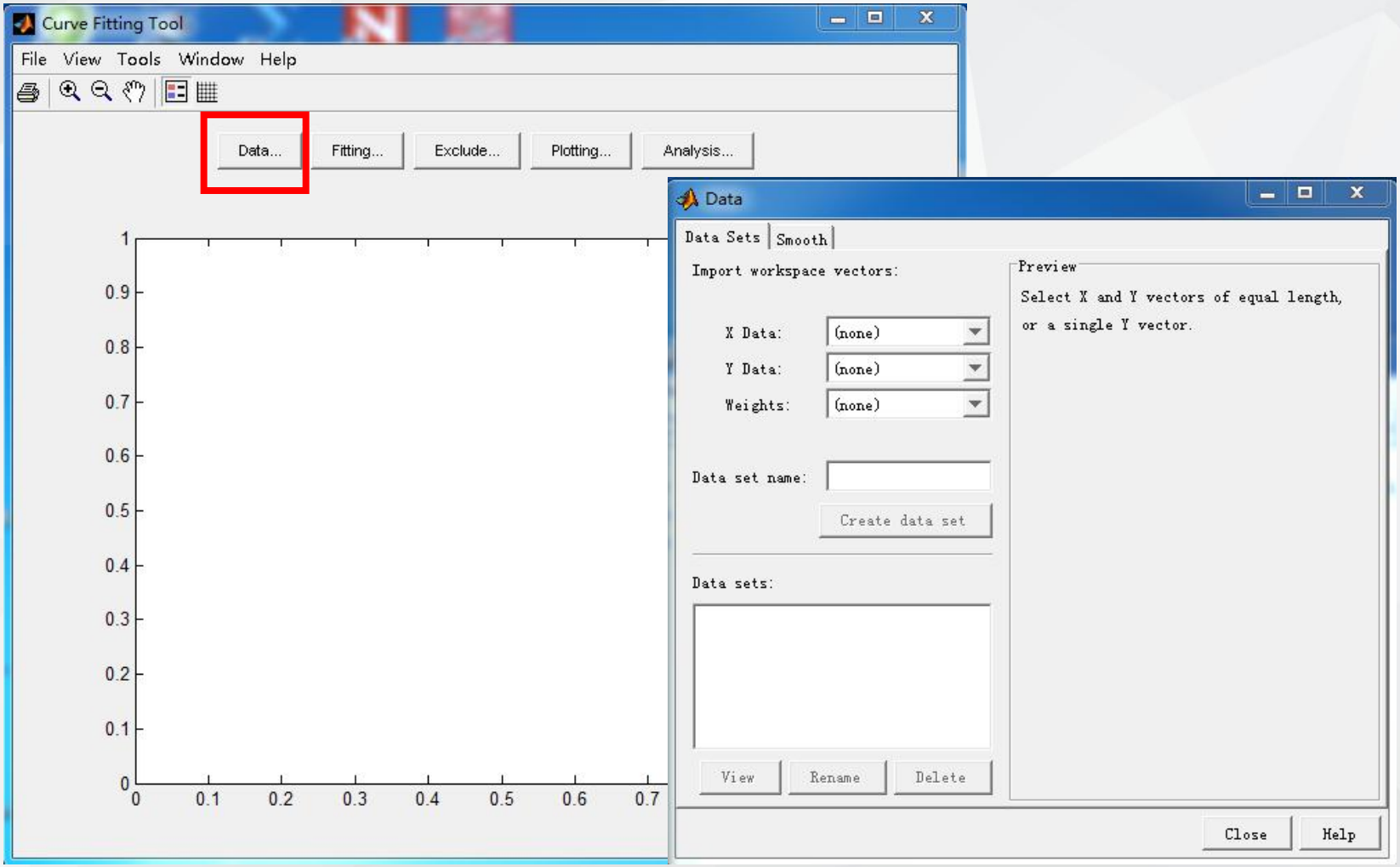
表4　氮肥的施肥量与土豆产量实验数据

| 氮肥量(kg) | 0 | 34 | 67 | 101 | 135 | 202 | 259 | 336 | 404 | 471 |
|---|---|---|---|---|---|---|---|---|---|---|
| 土豆产量(kg) | 15.18 | 21.36 | 25.72 | 32.29 | 34.03 | 39.45 | 43.15 | 43.46 | 40.83 | 30.75 |

请根据表4数据，分析土豆产量与氮肥施肥量之间的关系式。

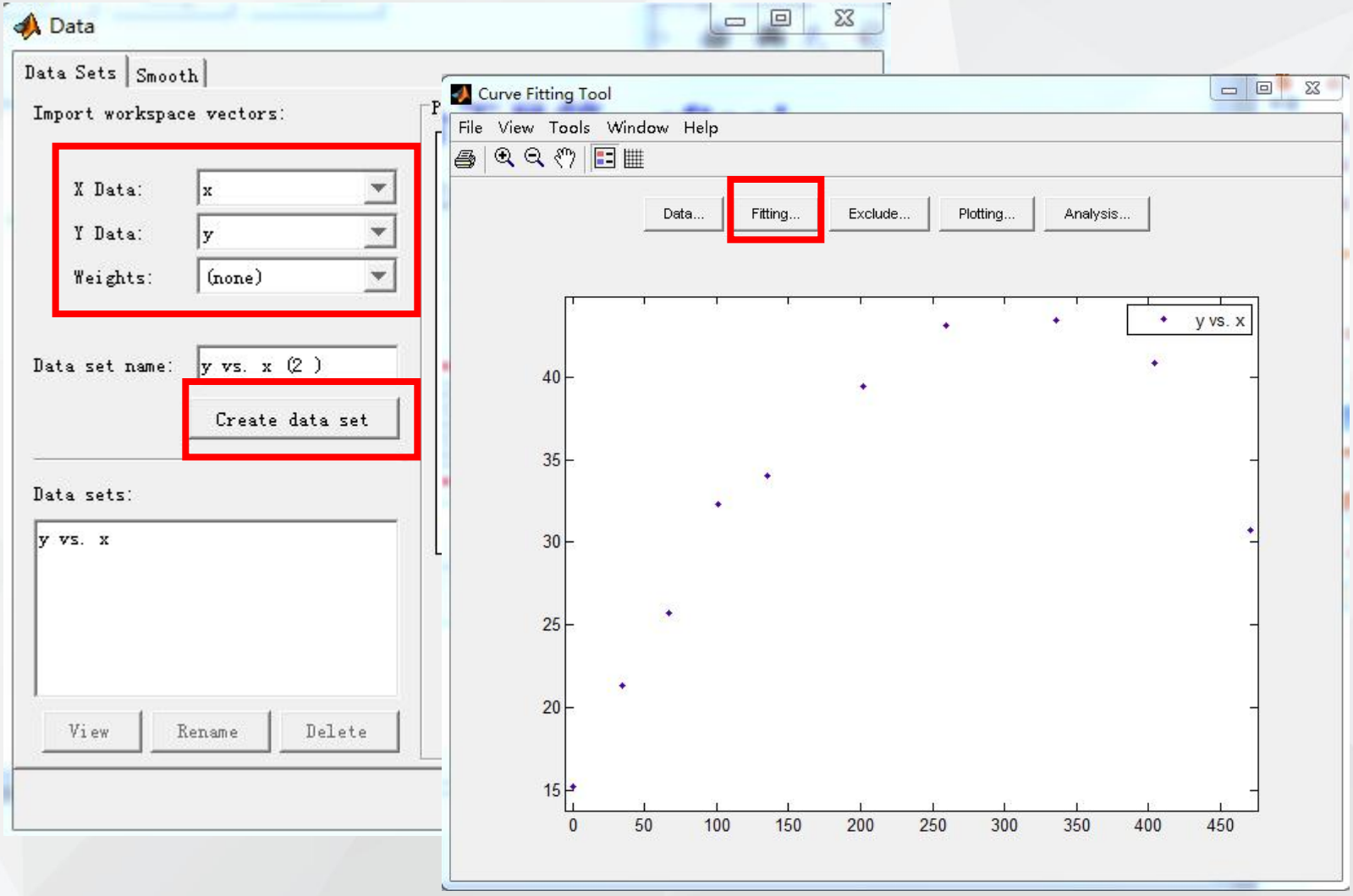
# Matlab拟合工具箱：cftool

在命令行窗口输入cftool，然后点回车打开拟合工具箱，单击DATA



# Matlab拟合工具箱：cftool

在红色框图内点下三角选择数据，x轴对应x的数据，y轴对应y的数据，然后单击close



# Matlab拟合工具箱：cftool

在红色框图内点下三角选择拟合类型：

Custom Equation-自定义公式，

Interpolant-插值逼近，

Lowess中的linear-线性拟合，

Lowess中的quadratic-二次方程组，

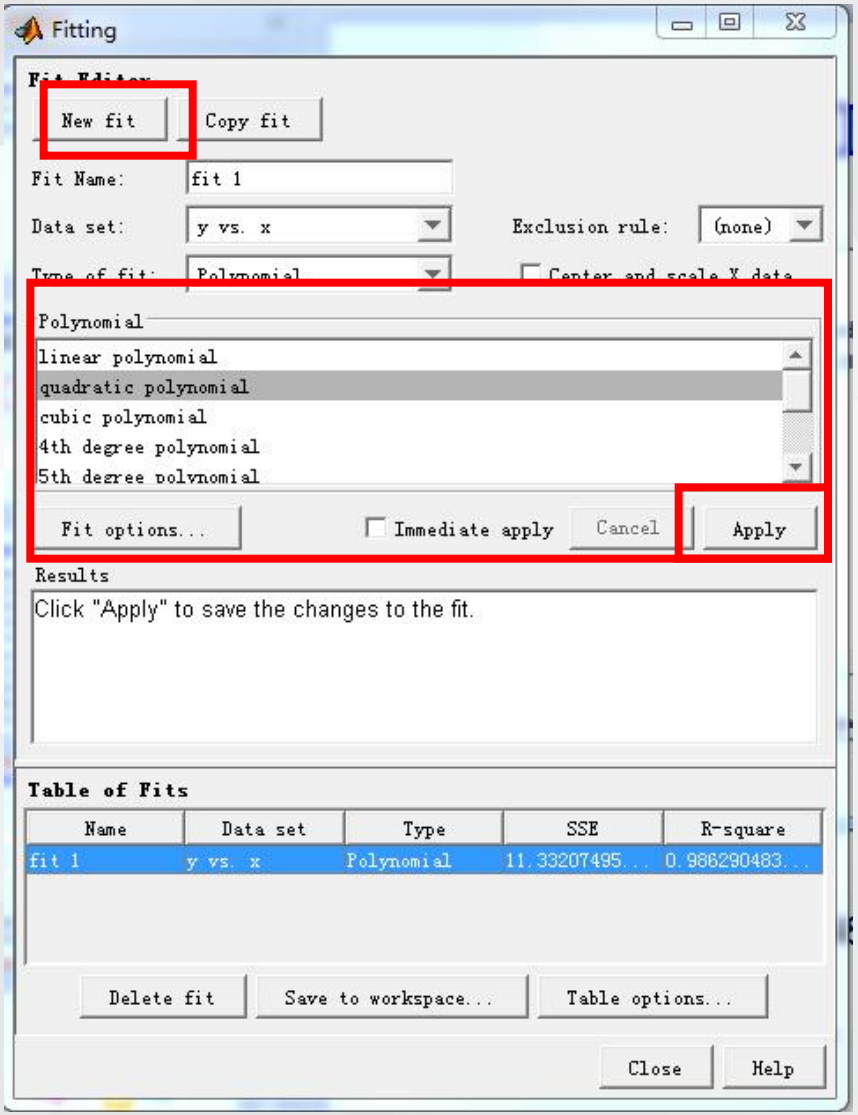
Polynomial-多项式逼近，

Power-幂函数逼近，

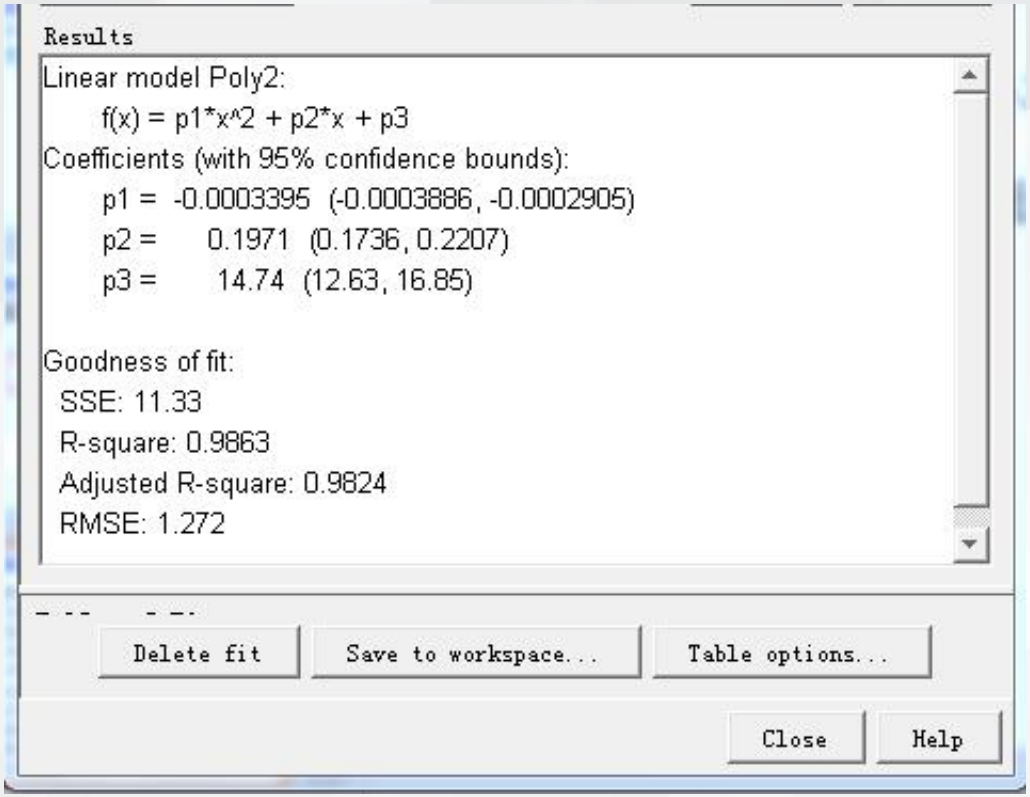
Gaussian-高斯逼近，

Expotential-指数逼近，

Fourier-傅里叶逼近



# Matlab拟合工具箱：cftool




拟合效果为:

SSE: 0.02135　误差平方和

R-square: 0.9985　确定系数

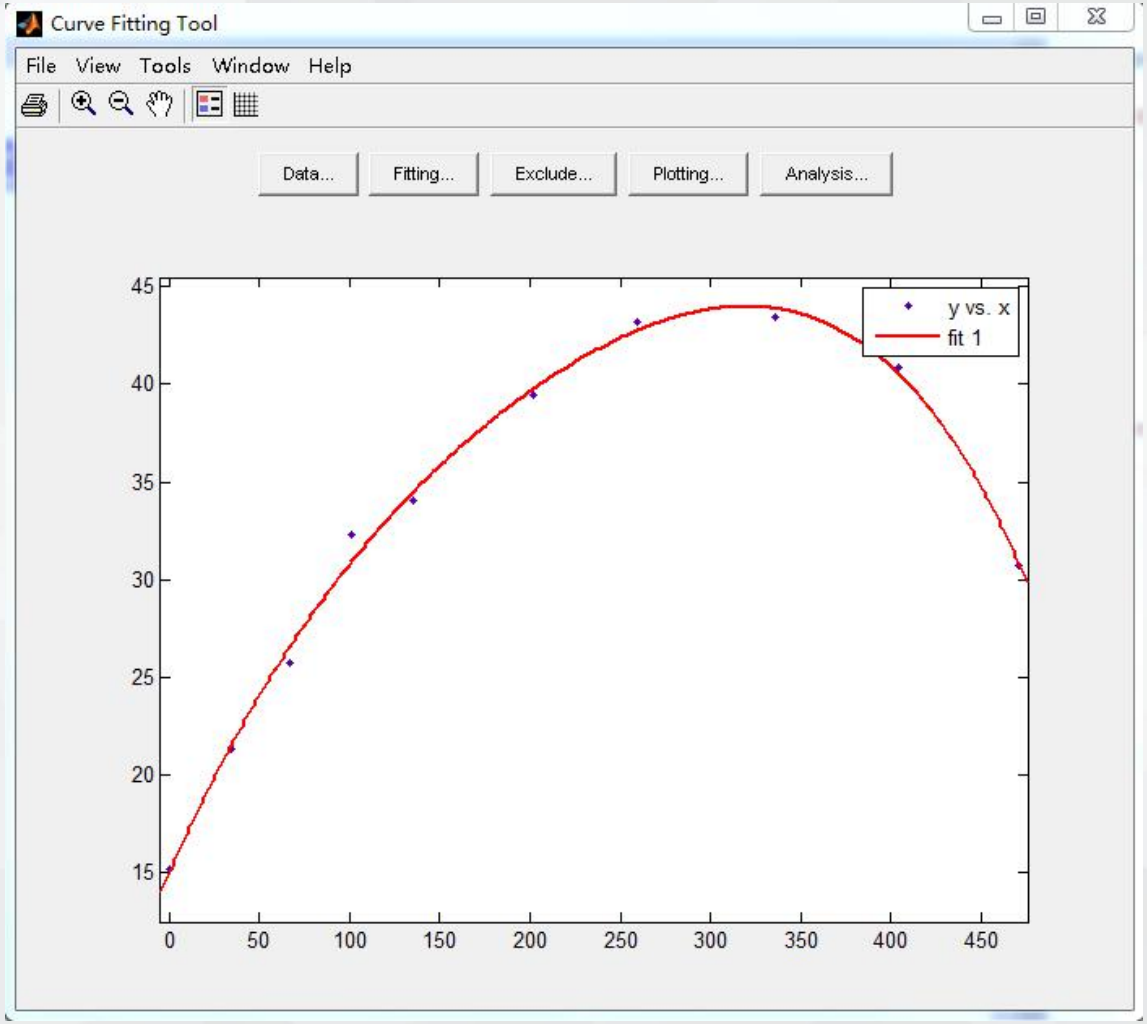
Adjusted R-square: 0.9983

RMSE: 0.03905　均方根差

拟合方程为y= -0.0003395*x^2 + 0.1971 *x + 14.74，

其中 **R-square越接近1，效果越好， SSE和RMSE越小越好。**

# Matlab拟合工具箱：cftool




观察拟合图形，点为对应的数据，线为我拟合出来的方程的图形


提
问
答
疑