



《数学实验》第11讲

主要内容:

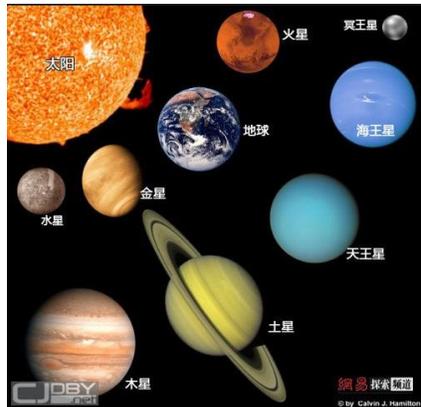
数据拟合

数据拟合之一——人口增长问题拟合

数据拟合之一——思考

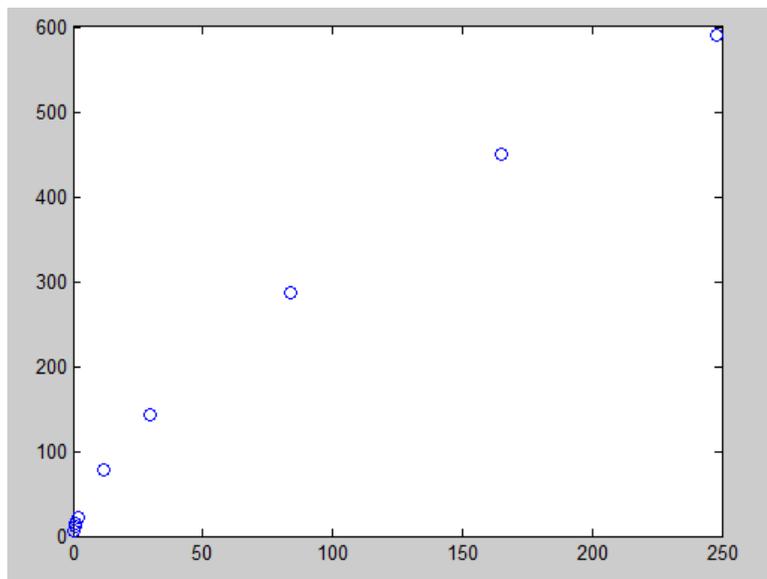
实践1：新冠肺炎人数预测

引例:



| 行星 | 轨道半 长轴 (10^{10} 米) | 周期 (年) |
|-----|-----------------------------|-----------|
| 水星 | 5.79 | 0.241 |
| 金星 | 10.8 | 0.615 |
| 地球 | 15.0 | 1 |
| 火星 | 22.8 | 1.88 |
| 木星 | 77.8 | 11.9 |
| 土星 | 143 | 29.5 |
| 天王星 | 287 | 84 |
| 海王星 | 450 | 165 |
| 冥王星 | 590 | 248 |

```
>> x=[0.241 0.615 1 1.88 11.9 29.5 84 165 248];  
>> y=[5.79 10.8 15.0 22.8 77.8 143 287 450 590];  
>> plot(x,y,'o')
```



开普勒大胆猜测：绕以太阳为焦点的椭圆轨道运行的所有行星，其各自椭圆轨道半长轴的立方与周期的平方之比是一个常量。即



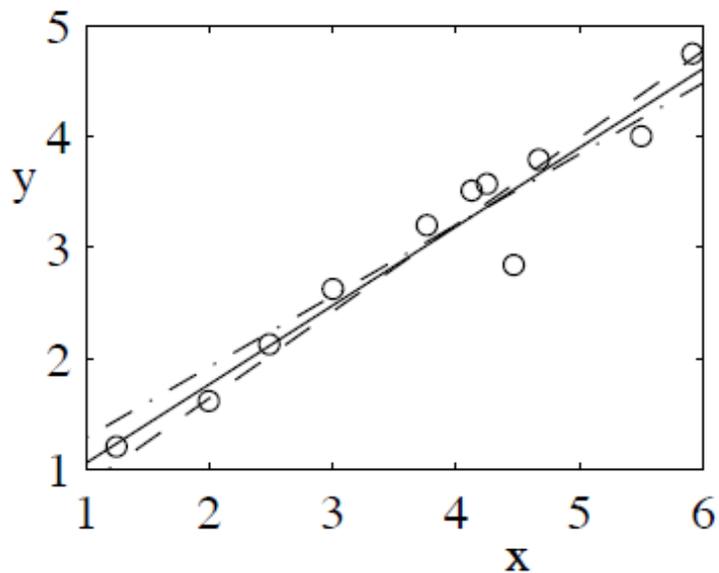
$$T^2 = ca^3$$

$$T^2 = ca^3$$

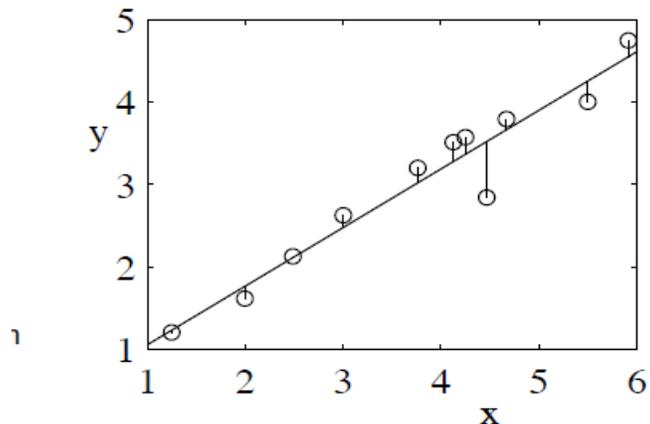
| 行星 | 半长轴 (a) | 周期 (T) | c |
|-----|-------------|------------|-----|
| 海王星 | 450 | 165 | 298 |
| 天王星 | 287 | 84 | 299 |

Q: 如何求 c ?

- ❖ 已知一组数据, 即平面上离散点 (x_i, y_i) , $i = 1, \dots, n$



- 寻求一条直线 $y=ax+b$ 使直线在某种规则下与所有数据点最为接近, 或称直线与离散数据拟合地最好。



离散点 (x_i, y_i) 与直线 $y=ax+b$ 的距离平方之和最小。即

$$S = (ax_1 + b - y_1)^2 + \cdots + (ax_n + b - y_n)^2$$

对 S 求 a 和 b 的偏导数，并令其为 0，得到如下方程组

$$(ax_1 + b - y_1) + \cdots + (ax_n + b - y_n) = 0$$

$$x_1(ax_1 + b - y_1) + \cdots + x_n(ax_n + b - y_n) = 0$$

整理得到：

$$\begin{pmatrix} 1 + \cdots + 1 & x_1 + \cdots + x_n \\ x_1 + \cdots + x_n & x_1^2 + \cdots + x_n^2 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} y_1 + \cdots + y_n \\ x_1 y_1 + \cdots + x_n y_n \end{pmatrix}$$

主要函数:

1. 多项式拟合函数: `polyfit()`

调用格式: `p=polyfit(x,y,n)`

输入: x, y —数据点, n —多项式阶数,

输出: p 为幂次从高到低的多项式系数向量

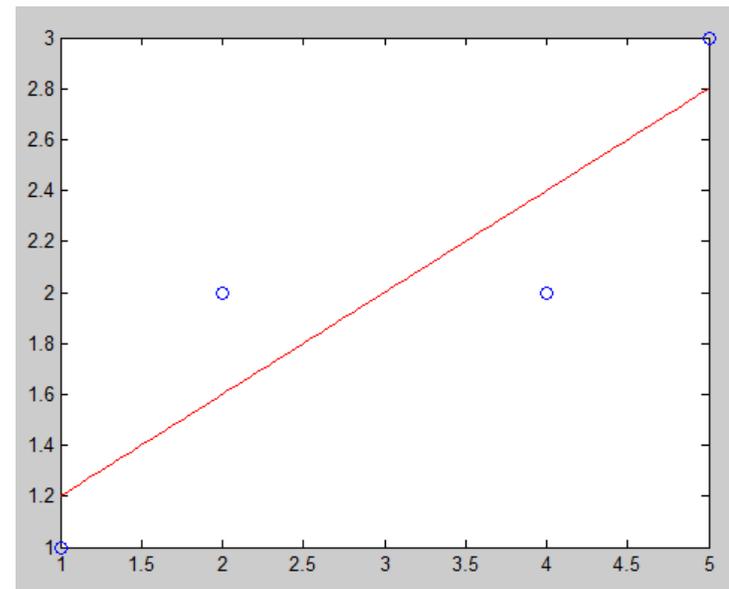
2. 多项式求值函数: `polyval()`

调用格式: `y=polyval(p,x)`

说明: `y=polyval(p,x)`为返回对应自变量 x 在给定系数 p 的多项式的值。

举例:

```
>> x = [1 2 4 5];  
>> y = [1 2 2 3];  
>> plot(x,y,'o')  
>> p=polyfit(x,y,1)  
>> xnew=linspace(min(x),max(x));  
>> ynew=polyval(p,xnew);  
>> hold on, plot(xnew,ynew, 'r')
```



一、数据拟合问题简介

数据拟合：从一大堆看上去杂乱无章的数据中找出规律性来，即设法构造一条曲线（拟合曲线）反映所给数据点总的趋势，以消除其局部波动。

常用拟合方法：多项式拟合

存在问题：并不是所有问题都可以用多项式作拟合，比如人口增长问题。

二、人口增长问题的数据拟合方法

问题：已知1996-2004年全国人口总数如下表，试根据表中数据预测2020年全国人口总数。（单位：亿）

| | | | | | |
|----|---------|---------|---------|---------|---------|
| 年 | 1996 | 1997 | 1998 | 1999 | 2000 |
| 人口 | 12.2389 | 12.3626 | 12.4761 | 12.5876 | 12.6743 |

| | | | |
|---------|---------|---------|---------|
| 2001 | 2002 | 2003 | 2004 |
| 12.7627 | 12.8453 | 12.9227 | 13.0000 |

分析：据人口增长的统计资料和人口理论数学模型知，当**人口总数N**不是很大时，在不太长的**时期t**内，人口增长接近于指数增长。

故采用指数函数对数据进行拟合

$$N = e^{a+bt}$$

为了计算方便，将上式两边同时取对数，得

$$\ln N = a + bt$$

令：

$$y = \ln(N) \quad N_i \rightarrow y_i$$

变换后的拟合函数为

$$y(t) = a + bt$$

(t_i, y_i)

由人口数据表对人口取对数，计算得：

| t | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| y | 2.5046 | 2.5147 | 2.5238 | 2.5327 | 2.5396 | 2.5465 | 2.5530 | 2.5590 | 2.5649 |

N_i

利用拟合函数polyfit对该组数据进行拟合，并计算当t=2020时的函数值：

```
clear,clc
```

```
t=[1996 1997 1998 1999 2000 2001 2002 2003 2004];
```

```
y=[2.5046 2.5147 2.5238 2.5327 2.5396 2.5465 2.5530 2.5590 2.5649];
```

```
p=polyfit(t,y,1)
```

```
Z=polyval(p,2020)
```

程序运行结果：

$p = 0.0074 \quad -12.3390$

$Z = 2.6864$

即:

$$a = -12.3390, \quad b = 0.0074$$

代入拟合函数得:

$$N = e^{-12.3390+0.0074t}$$

当 $t=2020$ 时, $N=14.6787$:

即到2020年时, 全国总人口数将达到14.6787亿

这一数据虽然不十分准确, 但是基本反映了人口变化趋势。

数据拟合——应用

- 1) 机器学习;
- 2) 数据预测;
- 3) 其他预测问题

机器学习:

① $(x_i, y_i), i=1, \dots, m$
 Input label

② $y = ax + b$

③ $\begin{cases} ax_1 + b = y_1 \\ \vdots \\ ax_m + b = y_m \end{cases}$

loss 数据

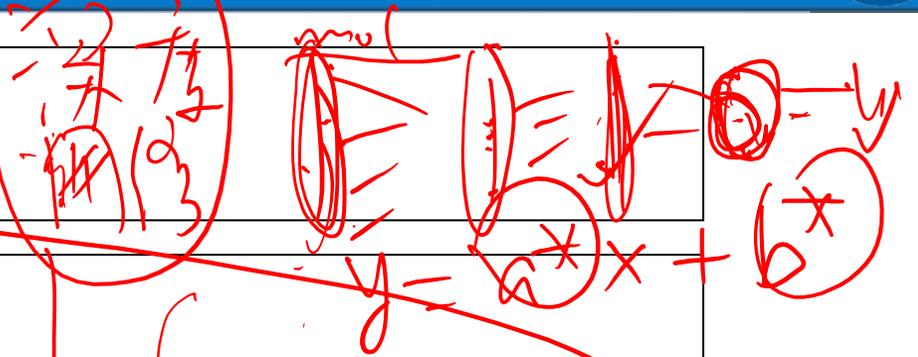
模型 + 1 测试
 参数

④ $\min_{a, b} \sum_{i=1}^m [y_i - (ax_i + b)]^2$ → output

⑤ $y = a^*x + b^*$

训练
 拟合
 参数

给定: x
 $y_1 = a^*x_1 + b^*$
 ↓ 预测值



实践1：新冠肺炎人数预测（20分钟）



新冠肺炎中国确诊人数/治愈人数预测及思考：

1) 数据获取（非常重要） <https://wh.opensource-service.cn/#/>

a) 请进入如上网站，手动收集随日期(存入 t)增加的肺炎确诊人数(存入 x)/治愈人数(存入 y)。搜集日期：从1月28日开始至2月28日；间隔：三天记录一次

举例：搜集1.28 – 2.06日数据，记录的 $t = [1, 4, 7, 10]$ ； $x = [4630, 9811, 17340, 28138]$ ； $y = [73, 214, 527, 1373]$

b) 请选用合适的多项式拟合你所收集的数据规律。

c) 根据上面的拟合结果，请预测1月30日、2月15日、2月21日、3月3日的确诊人数/治愈人数为多少？

d) 思考：i) 你的预测结果和实际数量相差大么？
ii) 如果相差大，问题在哪里/如何改进？

数据拟合

数据拟合之一——人口增长问题拟合

数据拟合之一——思考

实践1：新冠肺炎人数预测