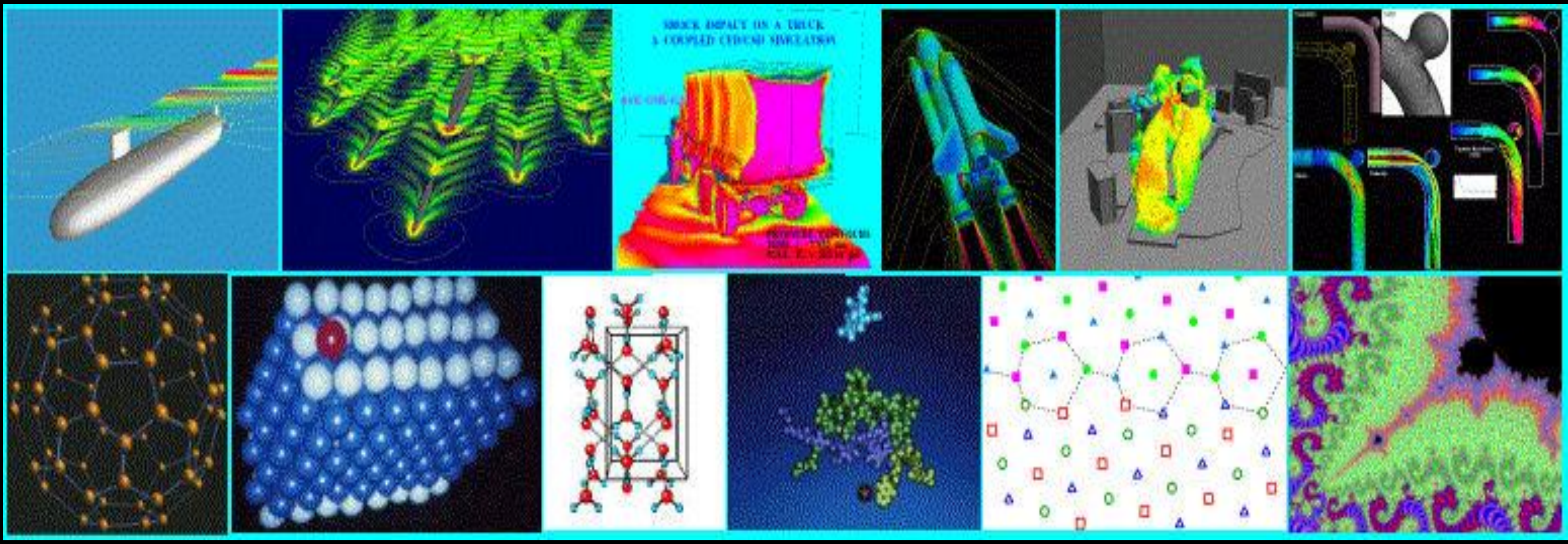
Computing for Scientists
Data Analysis (DA)
(April 23, 2013 – April 25, 2013)



Jie Zhang

CDS 130 - 003
Spring, 2013

Where We are?

Tool: MATLAB

0. Introduction & Syllabus

Section 1. Computer Fundamentals

Section 2. Scientific Simulation

Section 3. Visualization

Section 4: Data Analysis

Section 5: Ethics

We are here!

Section 4: Data Analysis (DA)

CH1. Introduction

CH2. Statistical Measures

CH3. Histogram method

CH4. Regression method

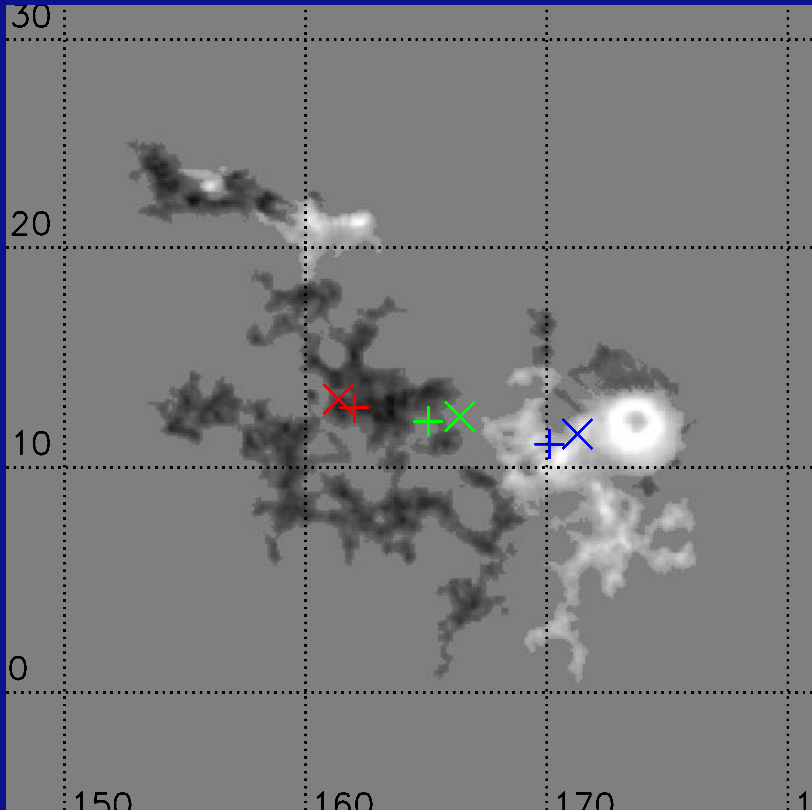
DA - CH1. Introduction
(April 23, 2013)

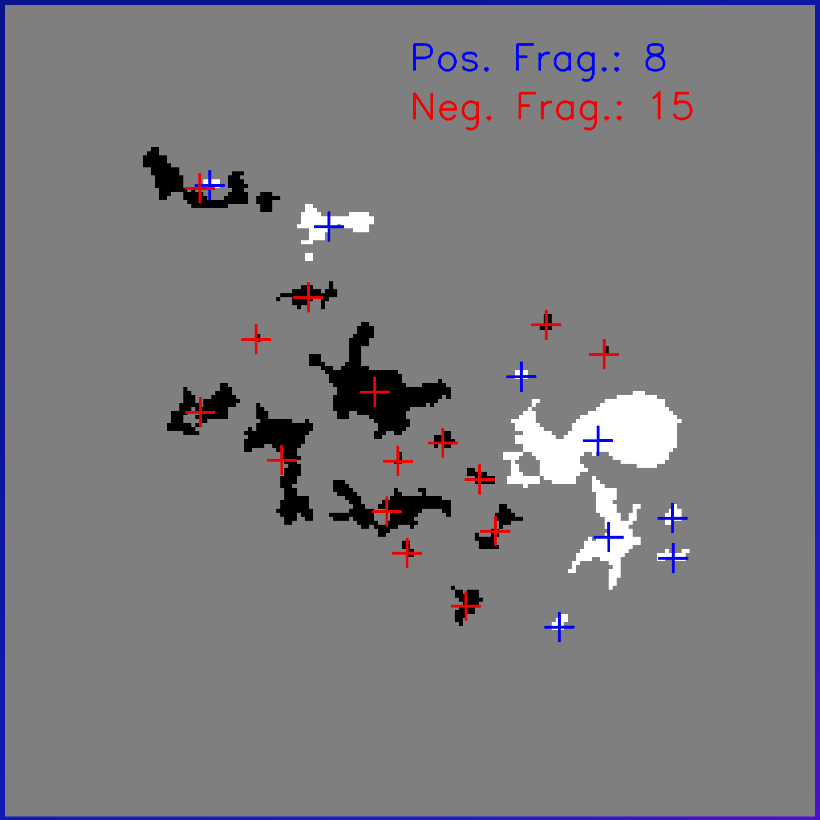
CH1. Introduction

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.

--http://en.wikipedia.org/wiki/Data_analysis

Exp. My Research on Sunspots





Find:
- **Coordinates**
- **Areas**
- **Fluxes**

Find:
- **Number of fragments ------ Analogous to number of sunspots.**

Matlab Overview of Data Analysis Tools

Watch video:

http://www.mathworks.com/videos/statistics-toolbox-overview-61211.html

DA – CH2. Statistical Measures

(April 23, 2013)

Statistical Measures

For a given data array, no matter 1-D, 2-D or 3-D data, one can always find:

- Minimum
- Maximum
- Median
- Mean
- Variance
- Standard Deviation

Minimum

Internal function for minimum is “min( )”

```
>a=[1,2,3;4,5,6;7,8,9]
a =
    1        2         3
    4        5         6
    7        8         9

>min(a)
ans   =
      1      2       3

>min(a(:))
ans   =  1


>a(:)         %what is this?
```

Why return 3 different numbers, not the minimum value of 1?

Answer: internal “min” function returns the minimum value of each column

“a(:)” converts the 2-D 3X3 array of “a” into 1-D data

Maximum

Internal function for maximum is “max( )”

```
>a=[1,2,3;4,5,6;7,8,9]
a =
    1       2       3
    4       5       6
    7       8       9

>max(a)
ans   =
      7      8      9

>max(a(:))
ans   =  9
```

Median

Internal function for median is “median( )”:
the median value is the mean of the middle two numbers in sorted order.

```
>a=[1,2,3;4,5,6;7,8,9]
a =
    1       2       3
    4       5       6
    7       8       9

>median(a)
ans   =
      4      5      6

>median(a(:))
ans   =  5

>median([1,2,3,4]     %?
```

mean

The mean is the mean value of a distribution

$$\text{mean} = \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + ...x_N$$

mean

Internal function for mean is “mean( )”

```
>a=[1,2,3;4,5,6;7,8,9]

%find the sum of values of the data array
>mysum=0
>for i=[1:9]
>mysum = mysum + a(i);
>end

>mymean=mysum/9
mymean=5

>mean(a(:))
ans  =  5

>sum(a(:))
```

Median versus Mean

```
>a=[1,2,3;4,5,6;7,8,100]    %having a skewed data
     point, or outlier in the distribution


>median(a(:))
ans = 5



>mean(a(:))
ans  =  15.111
```

Variance and Standard Deviation

Both are measures of how spread out a distribution is. In other words, they are measures of scattering of the data

Variance is the average squared deviation of each number from its mean

$$\text{variance} = \sigma^2 = \frac{\sum_{i}^{N} (x_i - \mu)^2}{N-1}$$

Standard deviation is the square root of variance

$$\text{Standard Deviation} = \sigma = \sqrt{\text{var}\,iance} = \sqrt{\frac{\sum_{i}^{N} (x_i - \mu)^2}{N-1}}$$

Variance and Standard Deviation

Question: Given the algorithm in the previous slide. Implement a MATLAB program to calculate the variance and standard deviation of the array a = [1, 2, 3, 4, 5, 6, 7, 8, 9]?

Variance and Standard Deviation

The code (refer to “statistic_1.m”)

```
a=[1,2,3;4,5,6;7,8,9];

%find the variance and standard deviation
MyMean = mean(a(:));
MyVar=0;

for i=[1:9]
    MyVar=MyVar+power((a(i)-MyMean),2);
end

MyVar=MyVar/8;
str=sprintf('My Variance = %d',MyVar);
disp(str);

MyDev=power(MyVar,0.5);
str=sprintf('My Deviation = %d', MyDev);
disp(str);
```

Variance and Standard Deviation

The Answer, i.e., the output of the program

```
My Variance = 7.500000e+00
My Deviation = 2.738613e+00
```

Variance and Standard Deviation

Internal function for variance is “var( )”
Internal function for standard deviation is “std( )”

```
>a=[1,2,3;4,5,6;7,8,9]

>var(a(:))
ans  = 7.5000

>std(a(:))
ans = 2.7386
```

Exercise 1

Exercise:

What are the statistical measures of the following 1-D data array? You can use the internal function for a quick calculation.
a=[1.2, 0.5, 0.9, 2.4, 1.8, 3.1, 2.2, 1.0]

The Answer:
Min=
max=
Mean=
Median=
Variance=
Standard Deviation=

Exercise

Exercise:

A daily temperature variation is stored as a 2-D data in an external ASCII file "temperature.dat". The first column is the time, and the second column is the temperature.

(1) Read the data into the Matlab
(2) Find the six statistical measures of the daily temperature variation
(3) Plot the variation of temperature versus the time

Exercise

“temperature.dat”

```
% Hourly temperature predicted in Fairfax VA (22030)
% on Nov. 10, 2010 (Wednesday) for 24 hours
% from 1 AM to midnight
1    45
2    44
3    43
4    42
5    41
6    40
7    40
8    43
9    46
10   49
11   53
12   55
13   56
14   57
15   57
16   56
17   53
18   49
19   46
20   44
21   43
22   42
23   41
24   41
```

Exercise

“dlmread”: internal Matalb File Input Function

- **Read ASCII-delimited file of numeric data into matrix**

•delimiter: e.g., comma ‘,’, space “, colon “:”
•R, C: specify the row and column where the upper left corner of the data lies in the file

```
>data=dlmread(‘temperature.dat’,’ ‘,3,0)  %read the data from the file

>time = data(:,1)     %obtain the time

>temp = data(:,2)   %obtain the temperature
```

Exercise

The Answer:

min = 40

Max = 57

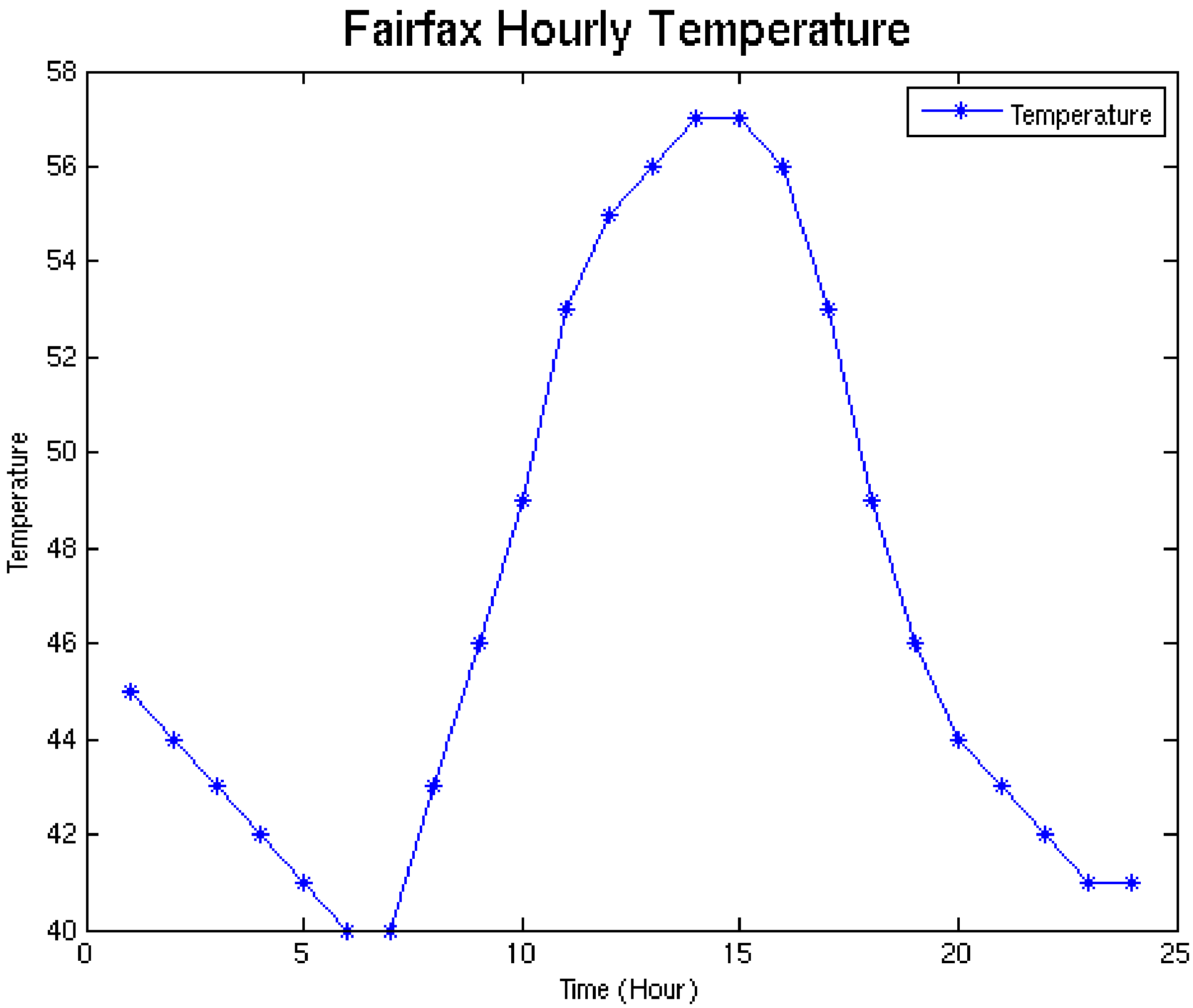
Median=44.5

Mean=46.91

Var=36.25

STD=6.02

Exercise



Fairfax Hourly Temperature
Temperature
58
56
54
52
50
48
46
44
42
40
0
5
10
15
20
25
Time (Hour)
Temperature

(April 23, 2013 Stopped Here)

April 25, 2013

Review: Statistical Measures

For a given data array, no matter 1-D, 2-D or 3-D data, one can always find:

- Minimum
- Maximum
- Median
- Mean
- Variance
- Standard Deviation

DA – CH3. Histogram

(April 25, 2013)

Histogram

Histogram is a summary graph showing the frequency distribution of data in various data range

- **Bin: the data range over which data points are counted, also called "groups"**
- **Frequency: number of data points on each bin**
- **Histogram can be shown in a bar plot, since each data bin is represented by one bar in the graph**

Histogram: Example

Raw data: “ASTR111_2007.dat” in an ASCII file

ASTR111-003, Astronomy, 2007, Instructor: Prof. Jie Zhang

ID	Grade
10000001	12
10000002	81
10000003	80
10000004	60
10000005	74
10000006	19
……………………	
……………………	
10000128	68
10000129	50
10000130	64
10000131	83
10000132	86

Histogram: Example

Refer to “histogram_1.m”

```
%column 1: student ID; column 2: student grade
%data starts from after the second column
%the delimiter is a space
a=dlmread('ASTR111_2007.dat',",2,0);

%obtain the grade from the second column
grad=a(:,2);

%specify the histogram bin
%the bin value indicates the center of the bin
bin=[5:10:95];

%count the frequency
%initialize the frequency distribution
freq(10)=0;
for i=[1:132]
   %bin 1: grade from 0 to 10
        if ((grad(i) >= 0) && (grad(i) < 10)), freq(1)=freq(1)+1; end
   %bin 2
        if ((grad(i) >= 10) && (grad(i) < 20)), freq(2)=freq(2)+1; end
        if ((grad(i) >= 20) && (grad(i) < 30)), freq(3)=freq(3)+1; end
        if ((grad(i) >= 30) && (grad(i) < 40)), freq(4)=freq(4)+1; end
        if ((grad(i) >= 40) && (grad(i) < 50)), freq(5)=freq(5)+1; end
        if ((grad(i) >= 50) && (grad(i) < 60)), freq(6)=freq(6)+1; end
        if ((grad(i) >= 60) && (grad(i) < 70)), freq(7)=freq(7)+1; end
        if ((grad(i) >= 70) && (grad(i) < 80)), freq(8)=freq(8)+1; end
        if ((grad(i) >= 80) && (grad(i) < 90)), freq(9)=freq(9)+1; end
        if ((grad(i) >= 90) && (grad(i) <= 100)), freq(10)=freq(10)+1; end
end

freq

plot(bin,freq,'-*b')
xlabel('bin','Fontsize',14)
ylabel('Frequency','Fontsize',14)
title('My Frequency Counter','FontSize',18)
legend('Astro Grade','Location','northwest')
```

- **Obtain the Data**
- **Count the frequency in data bin**

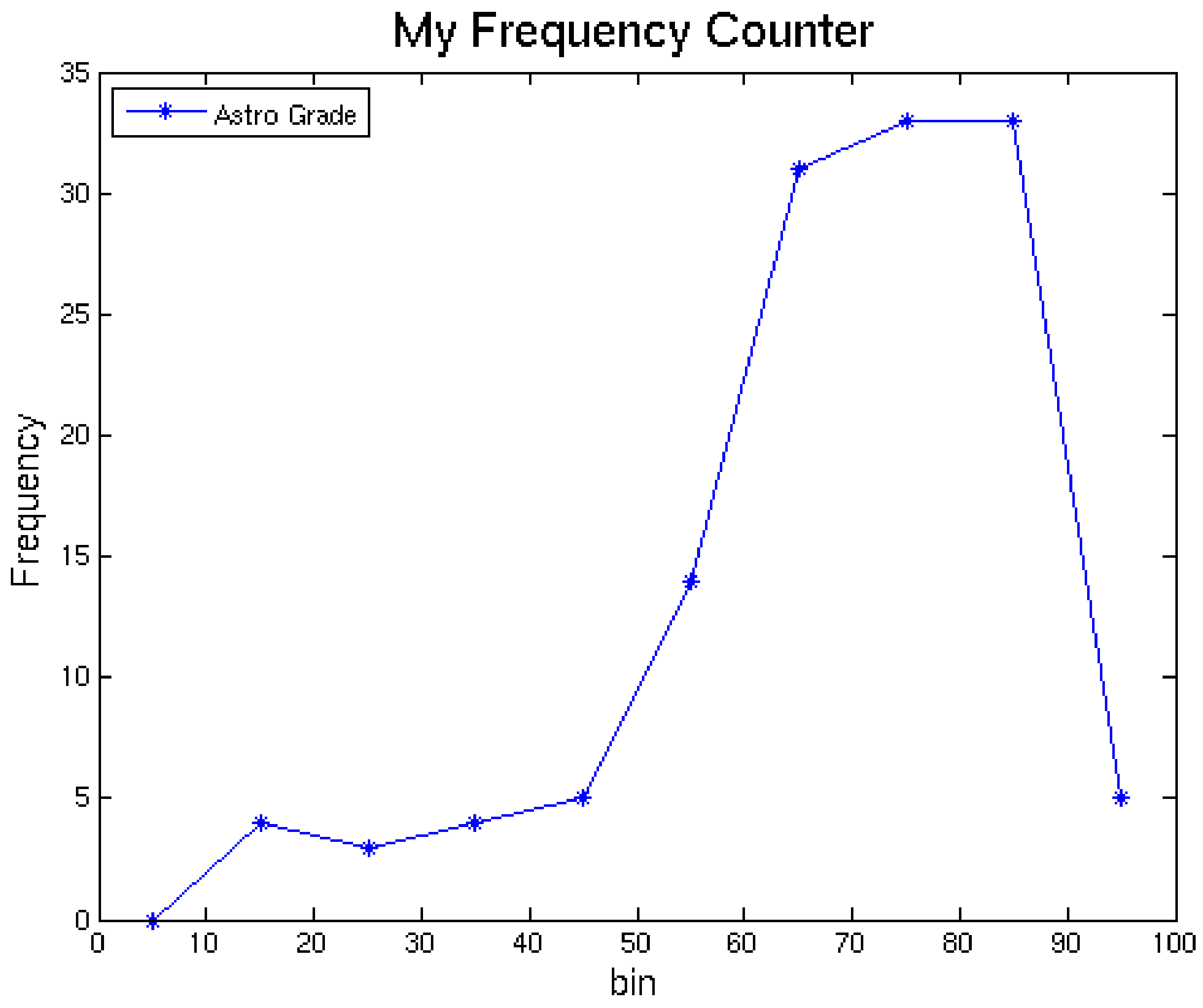
Histogram: Example

Frequency Table

Bin	Frequency
0-10	0
10-20	4
20-30	3
30-40	4
40-50	5
50-60	14
60-70	31
70-80	33
80-90	33
90-100	5

Histogram: Example

My own "poor" plot of the frequency distribution



“hist” method in Matlab

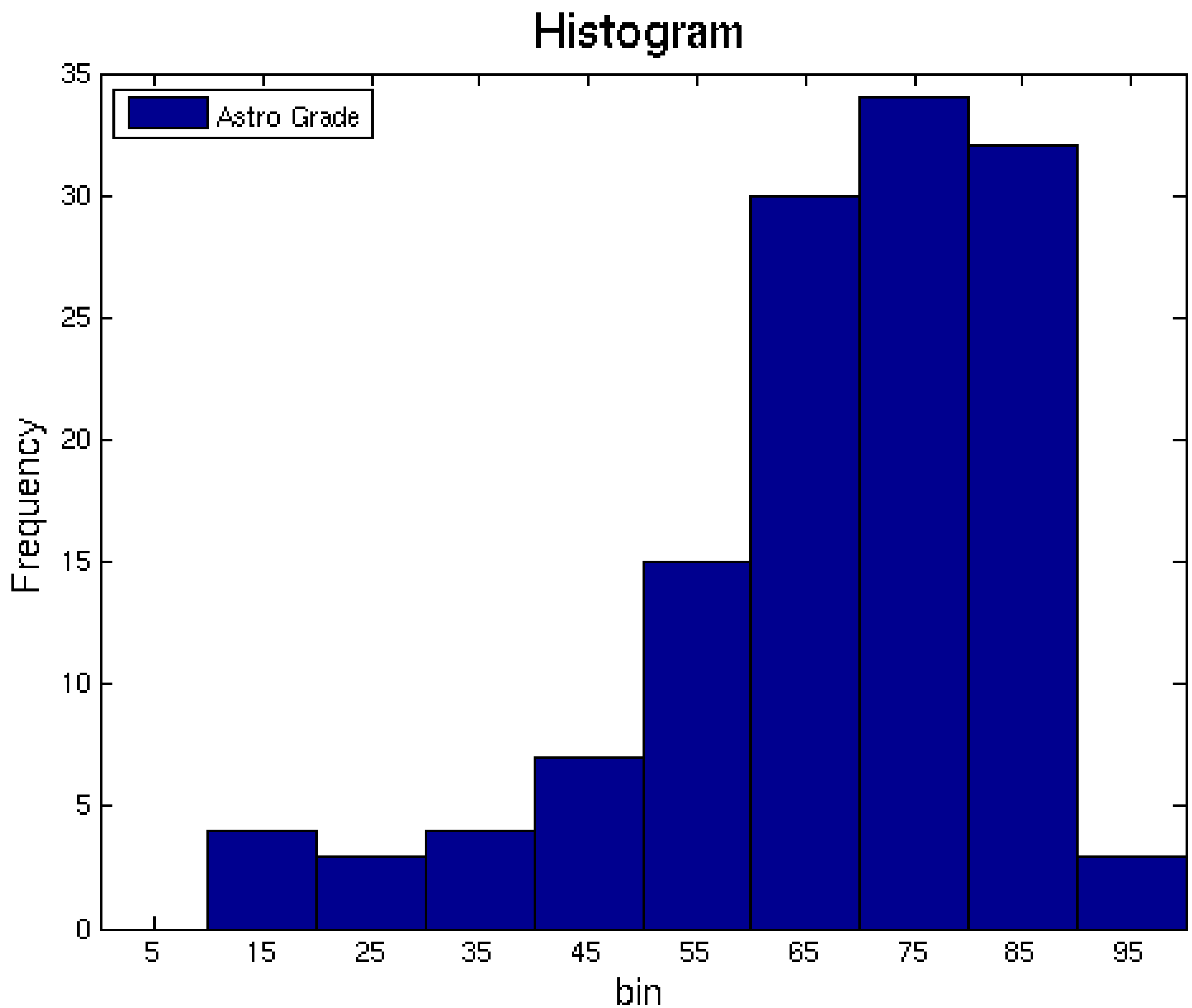
“hist” is a Matlab internal function

Refer to “histogram_2.m”

```
%read the data
data=dlmread('ASTR111_2007.dat','',2,0);

%obtain the grade from the second column
grade=data(:,2);

%default  bin: divide data into 10 bins
hist(grade)

%specify the bin; the value indicates the center of the bin
bin=[5:10:95]
hist(grade,bin)
%the return value of hist is the frequency value

xlabel('bin','Fontsize',14)
ylabel('Frequency','Fontsize',14)
title('Histogram','FontSize',18)
legend('Astro Grade','Location','northwest')
```

Histogram: Example

Output of “hist” method



Exercise

Generate 10000 random numbers from 0 to 1

>rand(10000,1)

(1) Plot histogram with 10 bins

>hist(rand(10000,1),10)

(2) How many numbers between [0.4,0.5]?

(3) Plot histogram with 50 bins

(4) Run “hist(rand(10,1),10)” multiple times, discuss the result

(5) Run “hist(rand(10000,1),10)” multiple times, discuss the result

DA – CH4.
Regression Method

(April 25, 2013)

Regression

Regression, or correlation, refers to the data analysis method to find the relationship that might exist between two scientific quantities

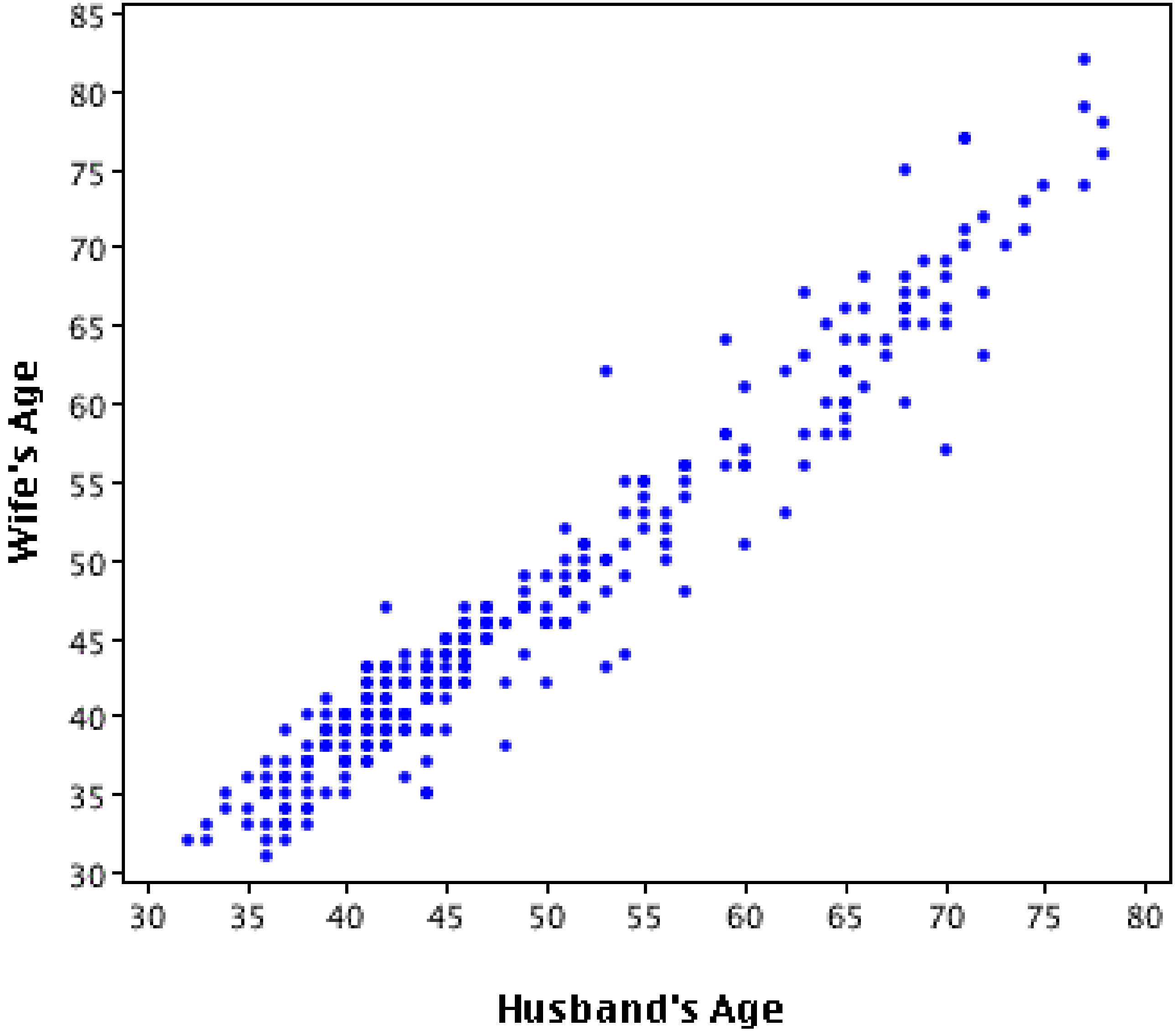
For example: one measures the height and weight for a group of people.

What are the data obtained? X (for height), Y (for weight)

What kind of useful analysis can you do with the data?

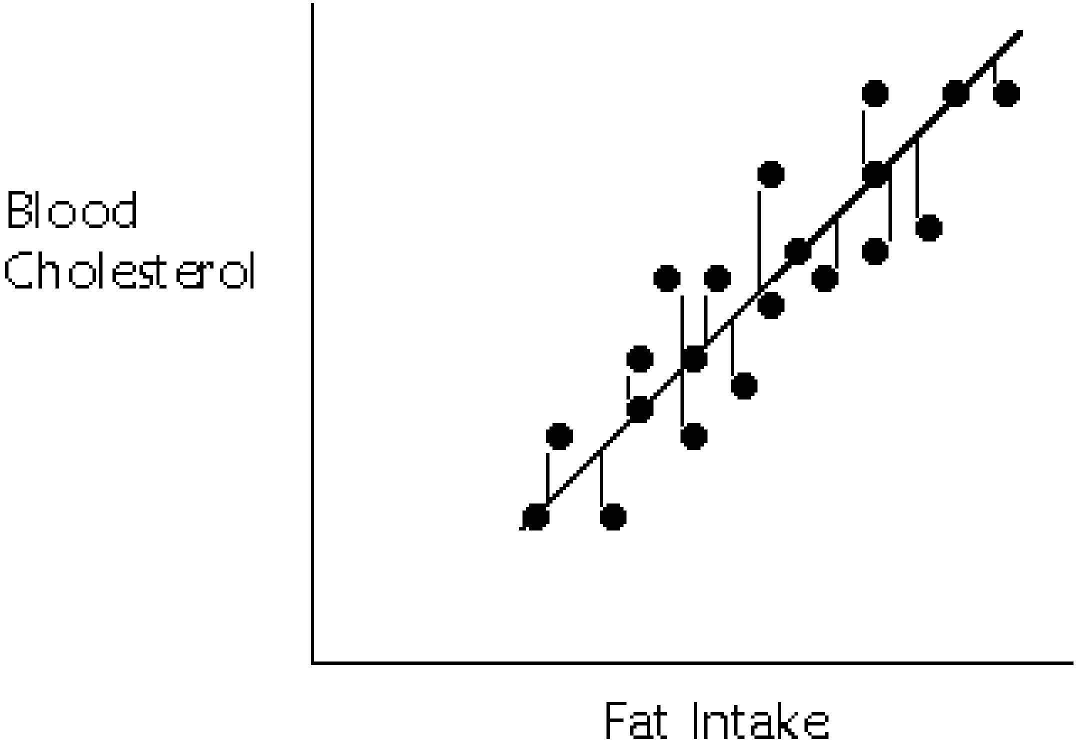
- **What can you do with the methods you've learned already?**
- **What else can you do?**

Example



85
80
75
70
65
60
55
50
45
40
35
30
Wife's Age
30 35 40 45 50 55 60 65 70 75 80
Husband's Age

Example



http://www.upa.pdx.edu/IOA/newsom/pa551/lectur19.htm

The Objectives

- The input of data is pair of quantities
  - x=[x1,x2,x3,x4,x5…]
  - y=[y1,y2,y3,y4,y5....]

- To quantify the relation of the two quantities
  - Objective one is to find the regression line that best fits the data -> the equation of the line
    - Useful for prediction: given X, find Y exactly
  - Objective two is to determine how well the line fits the data → correlation coefficient R

Example: Income and Education

Fifteen people are surveyed. The following table shows the data

Participants	Income ($)	Education (year)
1	125000	19
2	100000	20
3	90000	16
4	75000	16
5	100000	18
6	29000	12
7	35000	14
8	24000	12
9	50000	16
10	60000	17
11	30000	13
12	60000	15
13	56000	14
14	35000	12
15	90000	18

Example: Income and Education

data file in ASCII format, refer to "education_income.dat"

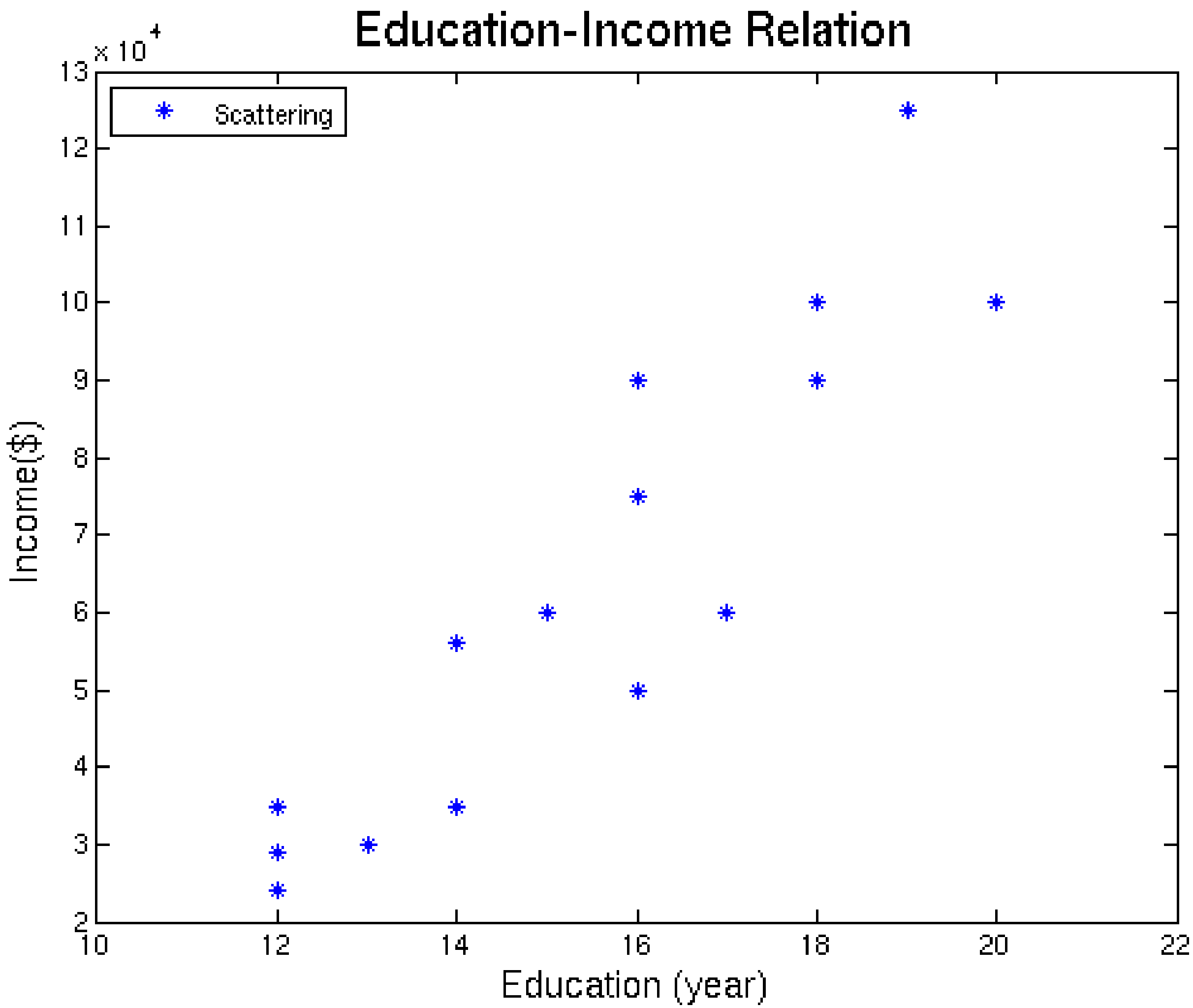
```
% education and income data
% survey of 15 people
%Three columns are: participants, Income and year of education
1   125000  19
2   100000  20
3    90000  16
4    75000  16
5   100000  18
6    29000  12
7    35000  14
8    24000  12
9    50000  16
10   60000  17
11   30000  13
12   60000  15
13   56000  14
14   35000  12
15   90000  18
```

Example: Income and Education

Read-in and plot the data, refer to “regression_edu_income_1.m”

```
%regression analysis 1
clear;clc
data=dlmread('education_income.dat','',3,0);
income=data(:,2);
edu=data(:,3);

%make scattering plot
plot(edu,income,'*b')                    % I omitted "-" here, why?
xlim([10,22])
ylim([20000,130000])
xlabel('Education (year)','Fontsize',14)
ylabel('Income($)','Fontsize',14)
title('Education-Income Relation','FontSize',18)
legend('Scattering','Location','northwest')
```

Example: Income and Education



What can you learn?
What are missing here?

Example: Income and Education

Full version, refer to "regression_edu_income_2.m"

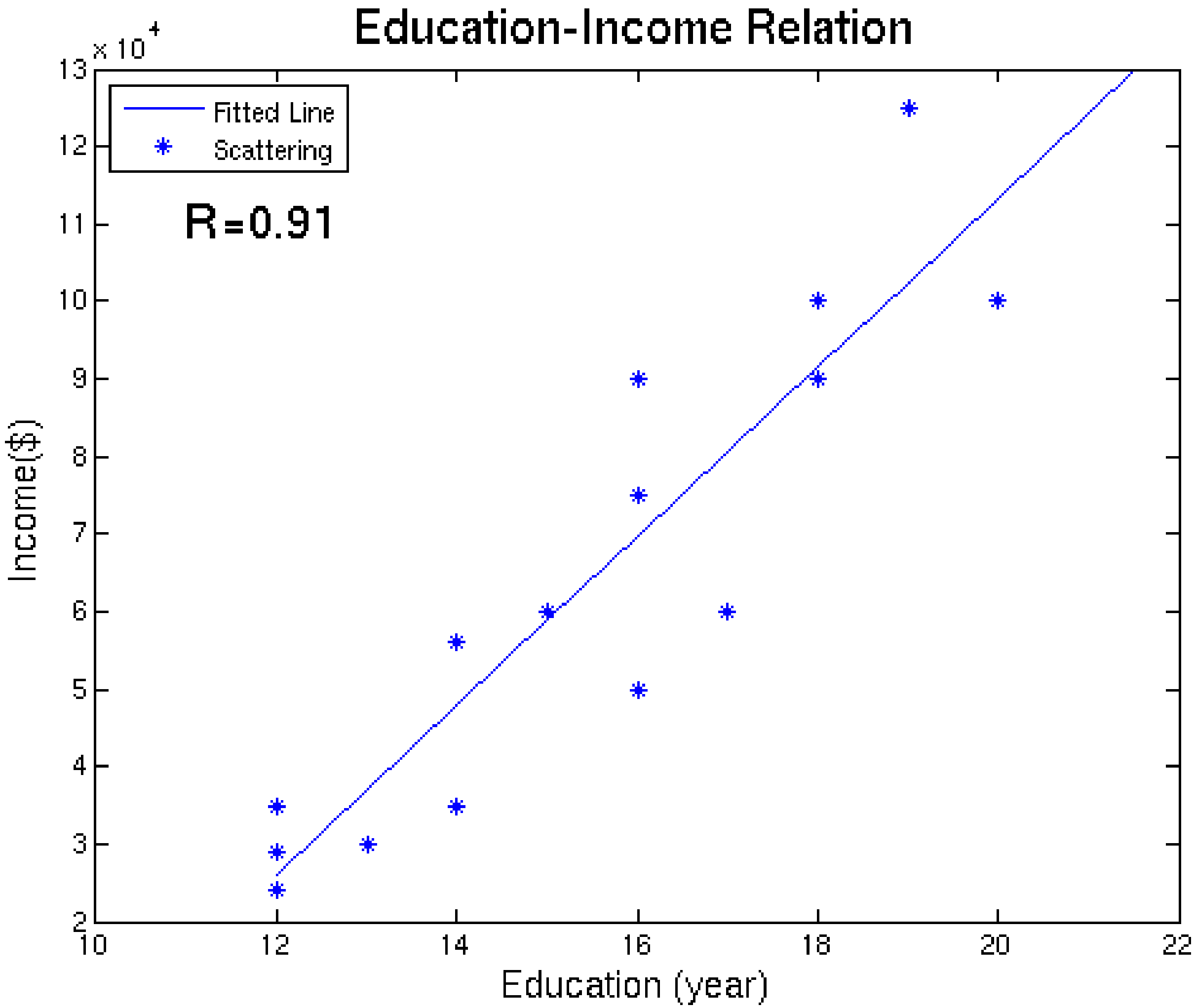
```
%regression analysis 2
clear;clc
data=dlmread('education_income.dat','',3,0);
income=data(:,2);
edu=data(:,3);

%make the linear regression
x=edu
y=income
p=polyfit(x,y,1)  %linear fit using polynomial method
%the fitted formula: y=p(1)*x + P(2)

%plot the line
x_fit=[12:22]   %define the x value of the fitted line
y_fit=p(1)*x_fit+p(2)  %predicted the Y-value using fitted function
plot(x_fit,y_fit,'-')

%obtain the correlation coefficient and annotate it
R=corrcoef(x,y)
text(11,110000,'R=0.91','Fontsize',20)

%make scattering plot
hold all
plot(x,y,'*b')
xlim([10,22])
ylim([20000,130000])
xlabel('Education (year)','Fontsize',14)
ylabel('Income($)','Fontsize',14)
title('Education-Income Relation','FontSize',18)
legend('Fitted Line','Scattering','Location','northwest')
```

Linear fitting to find the equation

Plot the fitting line

Find the correlation coefficient, and annotate it on the plot

Example: Income and Education



Linear Regression

Least square principle is used in the fitting

We fit the scattered data points into a linear equation

$$y = ax + b$$

$$\text{a : slope}$$

$$\text{b : y-interceptor}$$

Minimize the following "sum square" parameter to obtain the fitting coefficients "a" and "b"

$$SS = \sum_{i=1}^{N} (y_{fit}(i) - y_{obs}(i))^2$$

Matlab: “polyfit”

“polyfit.m”: internal Matlab function to make polynomial curve fitting to a pair of data point

```
>p=polyfit(x,y,n)    % nth order
>p=polyfit(x,y,1)    % n =1, linear fitting
```

$$p(x) = p_0 + p_1 x + p_2 x^2 + ... p_n x^n$$

n : degree of fitting

n : 1, linear fitting

n : 2, quadratic fitting

n : 3, cubic fitting

Linear Regression

```
>p = polyfit(x,y,1)   %x: x data
                      %y: y data
                      % degree of fitting, n=1 for linear
                      % p:fitting coefficient in descending power

p=
   1.0889e+004    -1.0449e+005

%p(1)=10889.: the slope, the "b" we are looking for
%p(2)=-104490: the y-interceptor, the "a" we are looking for
```

It means that the linear fitting function is:

$$y = ax + b$$

$$y = 10089.0x - 104490.0$$

Linear Regression: line plot

```
>x_fit=[12:22]      %define the x values of the fitted line
>y_fit=p(1)*x_fit+p(2)   %obtained the fitted y-value
>plot(x_fit,y_fit,'-')    %plot the line

>hold all              %add plot without erasing
```

Correlation Coefficient

Correlation coefficient, the R value, characterize how well the data is fitted by a linear function
R = 0, no correlation at all
R = 1.0, perfect correlation

$$R = \frac{ss_{xy}}{\sqrt{ss_{xx} ss_{yy}}}$$

$$ss_{xx} = \sum (x_i - \mu_x)^2$$

$$ss_{yy} = \sum (y_i - \mu_y)^2$$

$$ss_{xy} = \sum (x_i - \mu_x)(y_i - \mu_y)$$

Correlation Coefficient

“corrcoef.m”: the internal function of correlation coefficient

```
>R=corrcoef(x,y)
R = 0.90891
```

Correlation Coefficient

Annotating a text in the plot

```
> R=corrcoef(x,y)

> text(11,110000,'R=0.91','Fontsize',20)
```

Exercise

Make a linear regression analysis on the temperature-latitude data provided, refer to "temp_lat.dat"

(1) Plot the data points

(2) Fit the data points to a linear function, and find the function

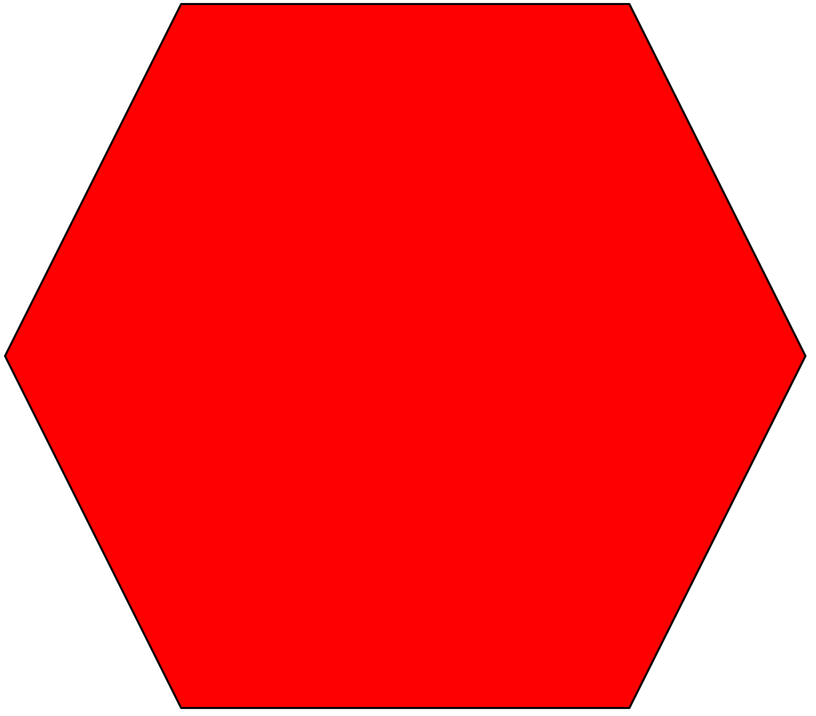
(3) Plot the fitted line in the same plot

(4) Find the correlation coefficient

(5) Annotate the correlation coefficient in the plot

Exercise

Latitude (degree)	**Temperature (Celcius)**
0	32
10	26
20	20
30	12
40	5
50	-1
60	-9
70	-14
80	-20
90	-25



The End