# Chapter 11 : data fitting using MATLAB

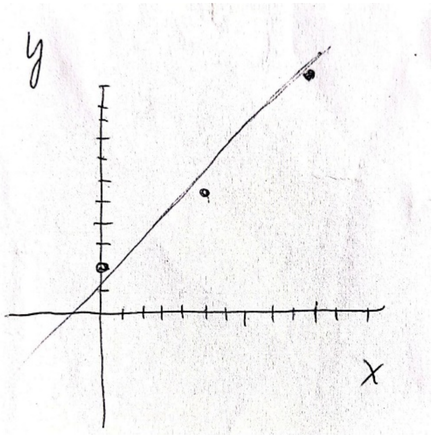
- Regression (Data fitting, Least square problem)
- Best fitting function
- (goodness of the fitting)
- Fit command details :
  - Centering & scaling
  - Creat fit options and fittype
- Interactive curve (data) & surface fitting

# The least square problem

**The Least-Squares Method**

Suppose we have the three data points given in the following table, and we need to determine the coef cients of the straight line $y = mx + b$ that best t the following data in the least-squares sense.

| $x$ | $y$ |
|---|---|
| 0 | 2 |
| 5 | 6 |
| 10 | 11 |




According to the least-squares criterion, the line that gives the best t is the one that minimizes $J$, the sum of the squares of the vertical differences between

the best fit line is the one that minimizes J,
: sum of the squares of the vertical
differences between the data points and the line.
called residuals.

The vertical differences between the line and the data points.

$$J = \sum_{i=1}^{3} (mx_i + b - y_i)^2$$

$$= (0m + b - 2)^2 + (5m + b - 6)^2 + (10m + b - 11)^2$$

The values of $m$ and $b$ that minimize $J$ are found by setting the partial derivatives $\partial J/\partial m$ and $\partial J/\partial b$ equal to zero.

$$\frac{\partial J}{\partial m} = 250m + 30b - 280 = 0$$

$$\frac{\partial J}{\partial b} = 30m + 6b - 38 = 0$$

m=0.9, b=11/6

Use linear, power, and exponential functions to describe data Each function will be a straight line when plotted for the axes specified in the following columns.

1. The linear function $y(x) = mx + b$ plotted on a linear axis results in a straight line. The slope is m and the intersection with the vertical axis is b.
2. The power function $y(x) = bx^m$ plotted on the full logarithmic axis resulting in a straight line.
3. The exponential function $y(x) = b(10)^{mx}$ or its equivalent $y(x) = b(e)^{mx}$

   Semi-logarithmic with the y-axis as the logarithm resulting in a straight line

# Processing steps to find the fitting function (linear, power, exponential) function.

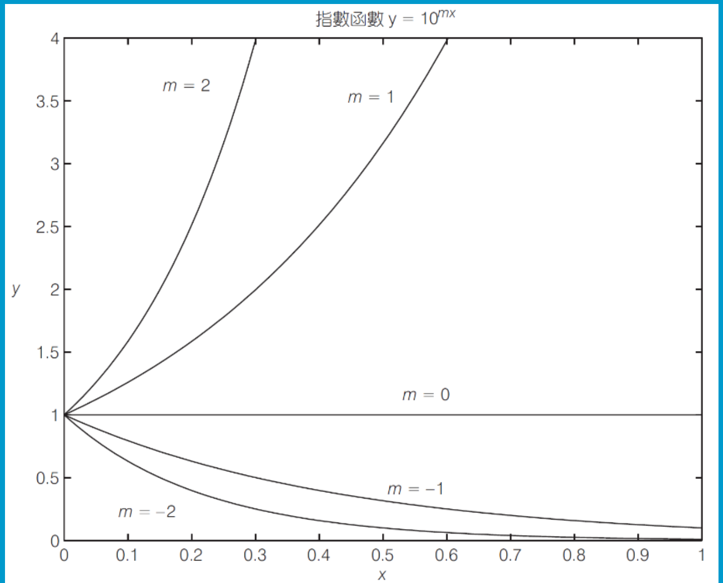
(1) Check the data close to the origin.

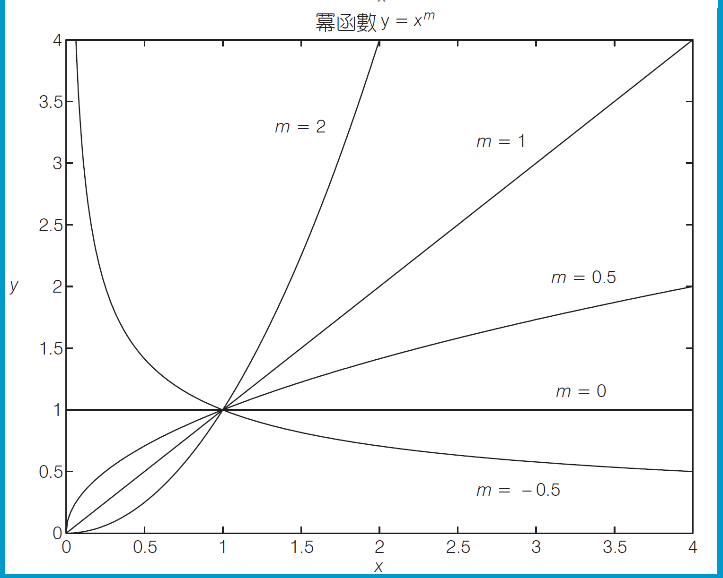
Exponential functions never go through the origin (unless b = 0, but that's meaningless).

(See the graph of the exponential function with b = 1 in Figure 6.1-1.)

A linear function passes through the origin with b = 0.

The power function only passes through the origin if m > 0. (See the graph of the power function with b = 1 in Figure 6.1-2.)



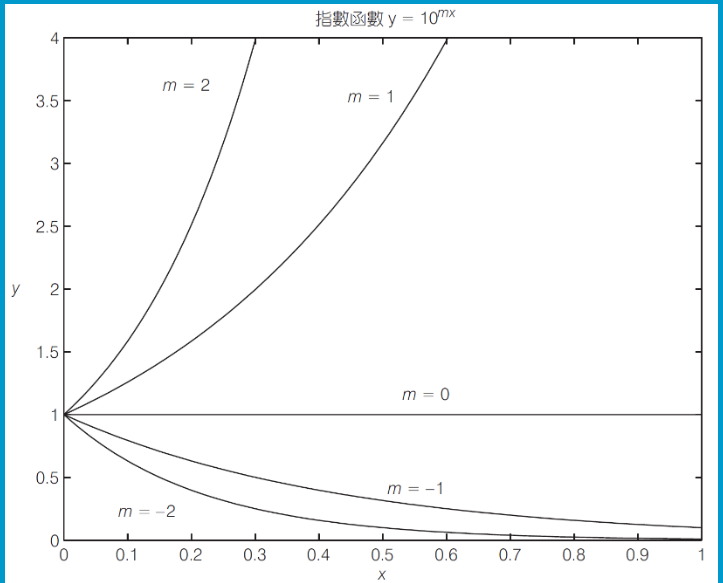



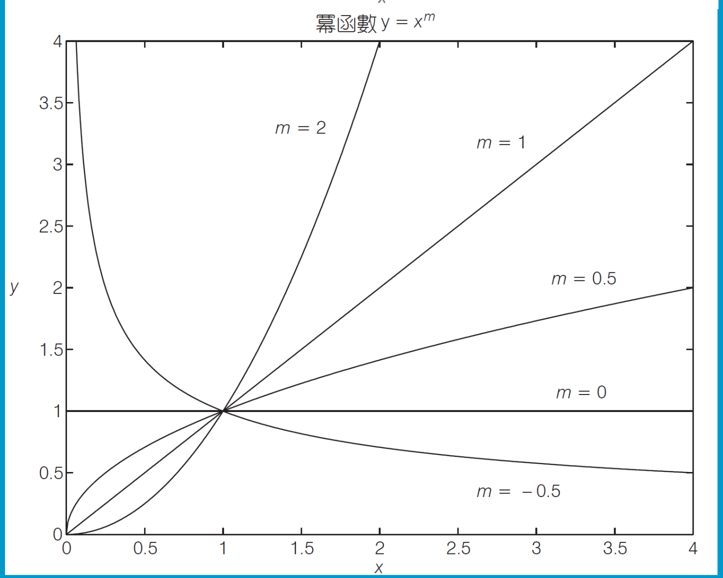
(2) Plot the data using a straight-line scale. If a straight line is obtained, it means that this data can be represented by a linear function, and the work is done. Conversely, if there is data at x = 0, then

a. If y(0) = 0, try using the power function.

b. If y(0) ≠ 0, try the exponential function.

If no data is given at x = 0, go to step 3.



3. If you suspect a power function, plot the data points on the full-log scale. -- a straight line on a full-logarithmic plot.

exponential function, plot the data on a semi-log scale. -- a straight line on a semi-logarithmic graph.

4. we use full-log or semi-log plots to identify function types, but do not obtain coefficients b and m, one use the polynomial fitting for power & exponential

# 圖 6.1–1　polyfit函數

| 指令 | 敘述 |
|---|---|
| `p = polyfit(x,y,n)` | 以 $n$ 次多項式擬合使用向量 $x$ 及 $y$ 所描述的資料，其中 $x$ 為自變數。傳回的向量 p 具有長度 $n+1$，所包含的元素是多項式的係數，並且以降冪排列。 |

# Use polyfit to fitting the data

Linear function: y = mx + b. In this case, the original data variables x and y, and find a fit by typing p = polyfit(x,y,1) Linear function.

P=[p1 p2] The first element p1 will be m, p2 will be b.

Power function:

$y(x) = bx^m$ . $\log_{10} y = m\log_{10} x + \log_{10} b$

Command p= polyfit( $\log_{10}$ (x), $\log_{10}$ (y),1),

p1 =m, p2 = $\log_{10}$ b. → b = $10^{p2}$

Exponential function:

$y(x) = b(10)^{mx}$. $\log_{10} y = mx + \log_{10} b$

Command p= polyfit(x, $\log_{10}$(y),1),

p1=m, p2 = $\log_{10}$ b. → b = $10^{p2}$

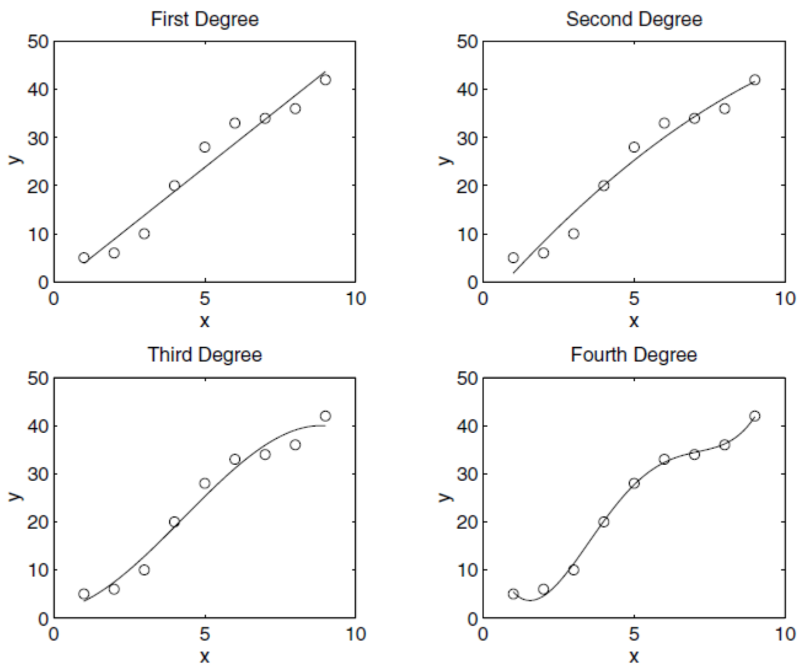
Two examples

Consider the data set where $x = 1, 2, 3, \ldots, 9$ and $y = 5, 6, 10, 20, 28, 33, 34, 36, 42$. The following script le computes the coef cients of the rst- through fourth-degree polynomials for these data and evaluates $J$ for each polynomial.

```
x = 1:9;
y = [5,6,10,20,28,33,34,36,42];
for k = 1:4
    coeff = polyfit(x,y,k)
    J(k) = sum((polyval(coeff,x)-y).^2)
end
```

The $J$ values are, to two signi cant gures, 72, 57, 42, and 4.7. Thus the value of $J$ decreases as the polynomial degree is ir

ure 6.2–1 shows this data and the four polyn



# Polynomial regression function

| Command | Description |
|---|---|
| p = polyfit(x,y,n) | Fits a polynomial of degree n to data described by the vectors x and y, where x is the independent variable. Returns a row vector p of length n+1 that contains the polynomial coef cients in order of descending powers. |
| [p,s, mu] = polyfit(x,y,n) | Fits a polynomial of degree n to data described by the vectors x and y, where x is the independent variable. Returns a row vector p of length n+1 that contains the polynomial coef cients in order of descending powers and a structure s for use with polyval to obtain error estimates for predictions. The optional output variable mu is a two-element vector containing the mean and standard deviation of x. |
| [y,delta] = polyval(p,x,s,mu) | Uses the optional output structure s generated by [p,s,mu] = poly t(x,y,n) to generate error estimates. If the errors in the data used with poly t are independent and normally distributed with constant variance, at least 50 percent of the data will lie within the band y ± delta. |

# how well the curve fits

We can use the J value to compare how well the curve fits two or more functions describing the same data.
The function that yields the smallest J value has the best fit to the data..

$$J = \sum_{i=1}^{m} [f(x_i) - y_i]^2$$

The sum of the squares of the amount of difference between our label value y and the mean $\bar{y}$ is S, which we can calculate by the following formula

$$S = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

$$r^2 = 1 - \frac{J}{S}$$

- For a perfect fit, J = 0 and $r^2 = 1$.
- Therefore, the closer $r^2$ is to 1,
- the better the fit $r^2$ is at most 1.
- J may be greater than S, so $r^2$ may be negative. If this happens, it means that this is a very bad model.
- As a rule of thumb, a good fit should account for at least 99% of the variance in the data. This value corresponds to $r^2 >= 0.99$.

# Three examples of the data fitting

- Goodness by using r square
- Use centering and scaling to improve numerical properties
- Create Fit option and Fit type before fitting
- Interactive curve (data) and surface fitting
- Exercise