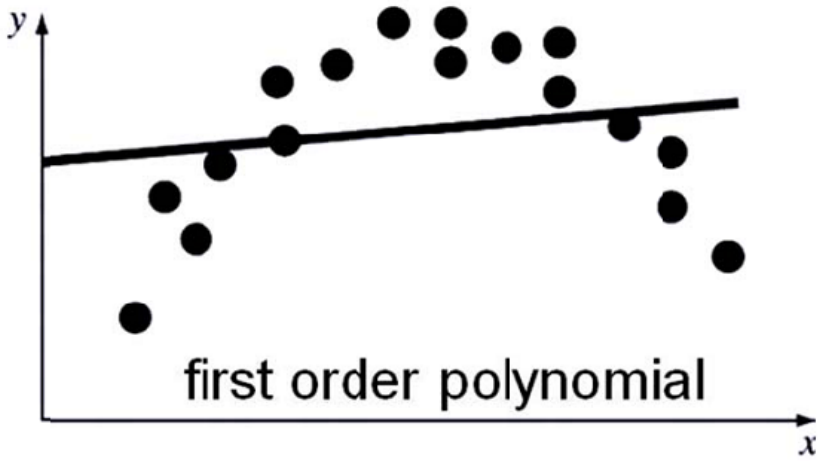
Second Midterm Exam: November 18 (Tuesday) 3:30pm – 5:30pm.
Lectures 6 – 20, Open-note, JH 245

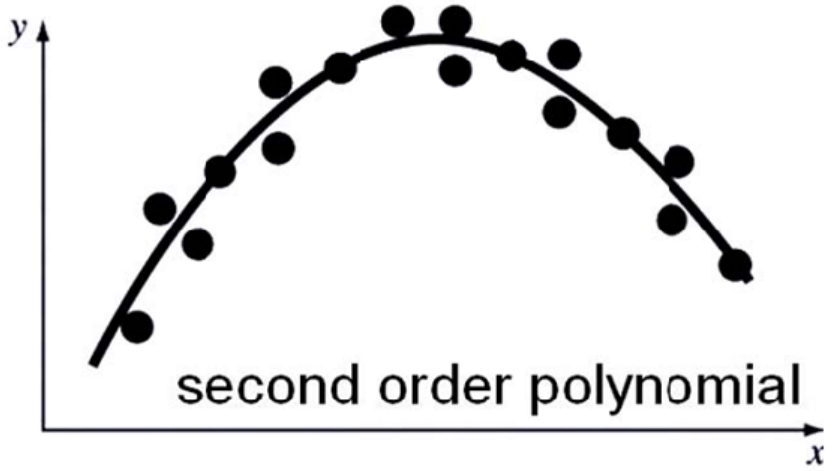
Lecture 22

Polynomial Regression

➢ **Polynomial Regression**

- The least-squares procedure can be readily extended to fit data to a higher-order polynomial. Again, the idea is to minimize the sum of the squares of the estimate residuals.





Polynomial Regression

➢ **Polynomial Regression**

- We have n data points in (x, y) space: (x_1, y_1), (x_2,y_2), … , and (x_n, y_n). We are going to fit these points to a polynomial.
- For a second order polynomial $y = a_0 + a_1x + a_2x^2$, the best fit for these data points would mean minimizing:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - y(x_i)]^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1x_i - a_2x_i^2\right)^2$$

y_i is the value of data points, $y(x_i)$ is the value estimated by the polynomial

- For a higher order polynomial $y = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m$, this would mean minimizing:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - y(x_i)]^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_mx^m\right)^2$$

- To get the coefficients of the polynomial, take the partial derivative of S_r with respect to a_i and make the derivatives zero. Then we get a system of linear equations with $m+1$ unknowns

- n is the number of data points, and m is the order of polynomial

Polynomial Regression

- **2nd Order Polynomial:** $S_r = \sum_{i=1}^{n}\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right)^2$

$$\frac{\partial S_r}{\partial a_0} = -2\sum_{i=1}^{n}\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{n} x_i\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum_{i=1}^{n} x_i^2\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right) = 0$$

These equations can be rearranged to normal linear equations,

$$(n)a_0 + (\Sigma x_i)a_1 + \left(\Sigma x_i^2\right)a_2 = \Sigma y_i$$

$$(\Sigma x_i)a_0 + \left(\Sigma x_i^2\right)a_1 + \left(\Sigma x_i^3\right)a_2 = \Sigma x_i y_i$$

$$\left(\Sigma x_i^2\right)a_0 + \left(\Sigma x_i^3\right)a_1 + \left(\Sigma x_i^4\right)a_2 = \Sigma x_i^2 y_i$$

The equations can be solved with the direct methods, Cramer's rule, or Gauss elimination methods.

Polynomial Regression

➢ m-th Order Polynomial

- The equations for two dimensional problems can be easily extended to m-th order polynomials, $y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - y(x_i)]^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_m x^m\right)^2$$

- To get the coefficients of the polynomial, take the partial derivative of S_r with respect to a_i and make the derivatives zero. Then we get a system of linear equations with $m+1$ unknowns

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_m x^m\right) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_m x^m\right) = 0$$

...

$$\frac{\partial S_r}{\partial a_m} = -2 \sum_{i=1}^{n} x_i^m \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_m x^m\right) = 0$$

Polynomial Regression

- **m-th Order Polynomial**
  - The equations can be rearranged to normal linear equations,

$$(n)a_0 + (\Sigma x_i)a_1 + \left(\Sigma x_i^2\right)a_2 + \cdots + (\Sigma x_i^m)a_m = \Sigma y_i$$

$$(\Sigma x_i)a_0 + \left(\Sigma x_i^2\right)a_1 + \left(\Sigma x_i^3\right)a_2 + \cdots + \left(\Sigma x_i^{m+1}\right)a_m = \Sigma x_i y_i$$

$$\left(\Sigma x_i^2\right)a_0 + \left(\Sigma x_i^3\right)a_1 + \left(\Sigma x_i^4\right)a_2 + \cdots + \left(\Sigma x_i^{m+2}\right)a_m = \Sigma x_i y_i$$

$$\ldots$$

$$(\Sigma x_i^m)a_0 + \left(\Sigma x_i^{m+1}\right)a_1 + \left(\Sigma x_i^{m+2}\right)a_2 + \cdots + \left(\Sigma x_i^{2m}\right)a_m = \Sigma x_i^m y_i$$

  - Summation Σ is done from i=1 to n.
  - The system of simultaneous linear equations can be solved with the direct methods, Cramer's rule, or Gauss elimination methods.

Linear Regression with More Independent Variables

➢ **Linear Regression with More Independent Variables**

- Another useful extension of linear regression is the case where y is a linear function of two or more independent variables, $x_1, x_2, \ldots, x_m$:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m$$

- We have n data points

$(x_{1,1}, x_{2,1}, \ldots, x_{m,1}, y_1)$ point 1

$(x_{1,2}, x_{2,2}, \ldots, x_{m,2}, y_2)$ point 2

…

$(x_{1,n}, x_{2,n}, \ldots, x_{m,n}, y_n)$ point n

The first subscript of x is the index of variable, the second the index of data points.
$x_{i,j}$ means the value of variable x_i at point j.

We are going to fit the linear function $y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m$ with these data points.

'n' is the number of data points, and 'm' is the number of independent variables.

Linear Regression with More Independent Variables

➢ **Two Variables**

- The best fit is obtained by minimizing the sum of the squares of the estimate residuals:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - y(x_{1,i}, x_{2,i})\right)^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}\right)^2$$

- To get the coefficients of the polynomial, take the partial derivative of S_r with respect to a_i and make the derivatives zero. Then we get a system of linear equations with m+1 unknowns

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}\right) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^{n} x_{1,i} \left(y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}\right) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum_{i=1}^{n} x_{2,i} \left(y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}\right) = 0$$

Linear Regression with More Independent Variables

Two Variables

- The equations can be rearranged to normal linear equations,

$$(n)a_0 + \left(\sum_{i=1}^{n} x_{1,i}\right)a_1 + \left(\sum_{i=1}^{n} x_{2,i}\right)a_2 = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_{1,i}\right)a_0 + \left(\sum_{i=1}^{n} x_{1,i}^2\right)a_1 + \left(\sum_{i=1}^{n} x_{1,i}x_{2,i}\right)a_2 = \sum_{i=1}^{n} x_{1,i}y_i$$

$$\left(\sum_{i=1}^{n} x_{2,i}\right)a_0 + \left(\sum_{i=1}^{n} x_{2,i}x_{1,i}\right)a_1 + \left(\sum_{i=1}^{n} x_{2,i}^2\right)a_2 = \sum_{i=1}^{n} x_{2,i}y_i$$

The system of simultaneous linear equations can be solved with the direct methods, Cramer's rule, or Gauss elimination methods.

Linear Regression with More Independent Variables

➢ **Linear Regression with More Independent Variables**

- The best fit is obtained by minimizing the sum of the squares of the estimate residuals:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_{1,i} - \ \ldots - a_m x_{m,i}\right)^2$$

- To get the coefficients of the polynomial, take the partial derivative of S_r with respect to a_i and make the derivatives zero. Then we get a system of linear equations with m+1 unknowns

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_{1,i} - \ \ldots - a_m x_{m,i}\right) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^{n} x_{1,i} \left(y_i - a_0 - a_1 x_{1,i} - \ \ldots - a_m x_{m,i}\right) = 0$$

$$\ldots$$

$$\frac{\partial S_r}{\partial a_m} = -2 \sum_{i=1}^{n} x_{m,i} \left(y_i - a_0 - a_1 x_{1,i} - \ \ldots - a_m x_{m,i}\right) = 0$$

Linear Regression with More Independent Variables

➢ Linear Regression with More Variables

- The equations can be rearranged to normal linear equations,

$$(n)a_0 + \left(\sum_{i=1}^{n} x_{1,i}\right)a_1 + \left(\sum_{i=1}^{n} x_{2,i}\right)a_2 + \cdots + \left(\sum_{i=1}^{n} x_{m,i}\right)a_m = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_{1,i}\right)a_0 + \left(\sum_{i=1}^{n} x_{1,i}^2\right)a_1 + \left(\sum_{i=1}^{n} x_{1,i}x_{2,i}\right)a_2 + \cdots + \left(\sum_{i=1}^{n} x_{1,i}x_{m,i}\right)a_m = \sum_{i=1}^{n} x_{1,i}y_i$$

$$\left(\sum_{i=1}^{n} x_{2,i}\right)a_0 + \left(\sum_{i=1}^{n} x_{2,i}x_{1,i}\right)a_1 + \left(\sum_{i=1}^{n} x_{2,i}^2\right)a_2 + \cdots + \left(\sum_{i=1}^{n} x_{2,i}x_{m,i}\right)a_m = \sum_{i=1}^{n} x_{2,i}y_i$$

...

$$\left(\sum_{i=1}^{n} x_{m,i}\right)a_0 + \left(\sum_{i=1}^{n} x_{m,i}x_{1,i}\right)a_1 + \left(\sum_{i=1}^{n} x_{m,i}x_{2,i}\right)a_2 + \cdots + \left(\sum_{i=1}^{n} x_{m,i}^2\right)a_m = \sum_{i=1}^{n} x_{m,i}y_i$$

The system of simultaneous linear equations can be solved with the direct methods, Cramer's rule, or Gauss elimination methods.

Linear Regression with More Independent Variables

➢ **Linea Regression with More Variables**

- Example, use the linear function $y = a_0 + a_1 x_1 + a_2 x_2$ to fit the following data points

x1	x2	Y
0	0	5
2	1	10
2.5	2	9
1	3	0
4	6	3
7	2	27

The data are stored in an external data file.

Linear Regression with More Independent Variables

- **More Variables**

```
clc
clear

load points.dat

C=points;

x1=C(:,1);
x2=C(:,2);
y=C(:,3);

n=length(x1);

X(1,1) = n;
X(1,2) = sum(x1);
X(1,3) = sum(x2);

X(2,1) = sum(x1);
X(2,2) = sum(x1.^2);
X(2,3) = sum(x1.*x2);

X(3,1) = sum(x2);
X(3,2) = sum(x2.*x1);
X(3,3) = sum(x2.^2);
```

X is the coefficient matrix of the linear equations

$$(n)a_0 + \left(\sum_{i=1}^{n} x_{1,i}\right)a_1 + \left(\sum_{i=1}^{n} x_{2,i}\right)a_2 = \Sigma y_i$$

$$\left(\sum_{i=1}^{n} x_{1,i}\right)a_0 + \left(\sum_{i=1}^{n} x_{1,i}^2\right)a_1 + \left(\sum_{i=1}^{n} x_{1,i}x_{2,i}\right)a_2 = \sum_{i=1}^{n} x_{1,i}y_i$$

$$\left(\sum_{i=1}^{n} x_{2,i}\right)a_0 + \left(\sum_{i=1}^{n} x_{2,i}x_{1,i}\right)a_1 + \left(\sum_{i=1}^{n} x_{2,i}^2\right)a_2 = \sum_{i=1}^{n} x_{2,i}y_i$$

```
B(1) = sum(y);
B(2) = sum(x1.*y);
B(3) = sum(x2.*y);

B = B';

%A = X\B;
A = inv(X)*B;

disp(A)
```

In the mathematical equations, the first subscript of x is the index of independent variables. In this problem, in matrix C, the index of independent variables is the column number. It is not always the case.

Linear Regression with More Independent Variables

- **Steps for Finding the Coefficients of the Function (**a0, a1, a2, …, am**)**

  Step 1: Find the coordinates of data points (x_{1i}, x_{2i}, …, x_{mi}, y_i)

  Step 2: Create the coefficient matrix according to Page 9 or 11

  Step 3: Solve the linear equations for a0, a1, a2, …, am.

For step 1. If the coordinates are stored as columns, then

x1 = C(:,1);

x2 = C(:,2);

y = C(:,3);

x1	x2	Y
0	0	5
2	1	10
2.5	2	9
1	3	0
4	6	3
7	2	27

For step 1. If the coordinates are stored as rows, then

x1 = C(1,:);

x2 = C(2,:);

y = C(3,:);

x1	0	2	2.5	1	4	7
x2	0	1	2	3	6	2
y	5	10	9	0	3	27

- **Homework on Canvas**