Module 5 – Polynomial Regression

Polynomial least squares regression

The general problem of approximating a set of data,

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

with an algebraic polynomial

$$P_m(x) = a_m x^m + a_{m-1} x^{m-1} \dots + a_1 x + a_0,$$

of degree $m < n - 1$, using the least squares procedure can be handled similarly as of linear least squares regression.

Let

$$y = P_m(x) + e \longrightarrow$$

$$e = (y - \hat{y})$$
$$= (y - P_m(x))$$

We choose the constants $a_0, a_1, \dots, a_m$ to minimize the sum of squares of errors. That is:

$$E = \sum_{i=1}^{n} \big(\underbrace{y_i - P_m(x_i)}\big)^2 = \sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} y_i P_m(x_i) + \sum_{i=1}^{n} \big(P_m(x_i)\big)^2$$

Or

$$E = \sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} \left[y_i \left(\sum_{k=1}^{m} a_k x_i^k \right) \right] + \sum_{i=1}^{n} \left(\sum_{k=1}^{m} a_k x_i^k \right)^2$$

For E to be minimized, we need to have the following systems of linear equations with respect to unknowns solved:

$$\frac{\partial E}{\partial a_j} = 0, \qquad j = 0,1,2, \dots , m$$

This leads to the following system of linear equations with $m + 1$ equations and $m + 1$ unknows (***normal equations***)

$$a_0 \sum_{i=1}^{n} x_i^0 + a_1 \sum_{i=1}^{n} x_i^1 + a_2 \sum_{i=1}^{n} x_i^2 + \cdots + a_m \sum_{i=1}^{n} x_i^m = \sum_{i=1}^{n} y_i x_i^0,$$

$$a_0 \sum_{i=1}^{n} x_i^1 + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 + \cdots + a_m \sum_{i=1}^{n} x_i^{m+1} = \sum_{i=1}^{n} y_i x_i^1,$$

$$\vdots$$

$$a_0 \sum_{i=1}^{n} x_i^m + a_1 \sum_{i=1}^{n} x_i^{m+1} + a_2 \sum_{i=1}^{n} x_i^{m+2} + \cdots + a_m \sum_{i=1}^{n} x_i^{2m} = \sum_{i=1}^{n} y_i x_i^m.$$

$a_0 \sum_{i=1}^{n} x_i^0$

$a_0 \sum_{i=1}^{n} 1$

These *normal equations* have a **unique solution provided that the x_i are distinct**.

$m=2 \mapsto$ # of Variables = 3

# of eqns = 3

Example. Fit the data in the following table with the discrete least squares polynomial of degree at most 2.

i	x_i	y_i
1	0	1.0000
2	0.25	1.2840
3	0.50	1.6487
4	0.75	2.1170
5	1.00	2.7183

$$\hat{y} = a_0 + a_1 x + a_2 x^2 = P_2(x)$$

We use only 2-decimal places:

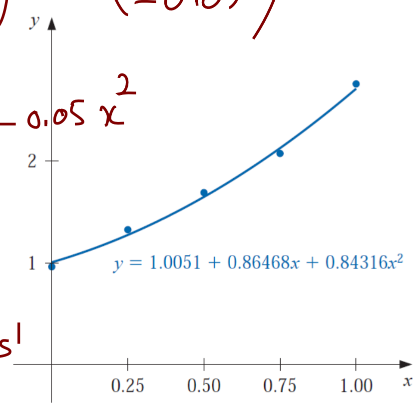
	x_i	y_i	x_i^2	x_i^3	x_i^4	$y_i x_i$	$y_i x_i^2$
	0	1	0	0	0	0	0
	0.25	1.28	0.06	0.02	0	0.32	0.08
	0.5	1.65	0.25	0.13	0.06	0.83	0.41
	0.75	2.12	0.56	0.42	0.32	1.61	1.19
	1	2.72	1	1	1	2.72	2.72
Sum =	2.5	8.77	1.87	1.57	1.38	5.48	4.4

$$\begin{cases} 5a_0 + 2.5a_1 + 1.87a_2 = 8.77 \\ 2.5a_0 + 1.87a_1 + 1.57a_2 = 5.48 \\ 1.87a_0 + 1.57a_1 + 1.38a_2 = 4.4 \end{cases} \longmapsto \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0.86 \\ 1.82 \\ -0.05 \end{pmatrix}$$

$$\begin{bmatrix} 5 & 2.5 & 1.87 \\ 2.5 & 1.87 & 1.57 \\ 1.87 & 1.57 & 1.38 \end{bmatrix}$$

$$\hat{y} = 0.86 + 1.82x - 0.05x^2$$

double-check all computations!



Note that the total error of this procedure will be

$$E = \sum_{i=1}^{5} (y_i - 1.0051 - 0.86468x_i - 0.84316x_i^2)^2$$

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

General Linear Least Squares Regression.

observations: $(x_1, x_2, \ldots, x_m, y)$

The idea of linear least squares regression can be extended to the case we have more than one independent variable. Assume that y is related to the independent variables $x_1, x_2, \ldots, x_m$ using the following linear form:

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

$e_i = y_i - \hat{y}_i$

As before, the "best" values of the coefficients are determined by formulating the sum of the squares of the residuals:

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

where $(x_{1,i}, x_{2,i}, \ldots, x_{m,i}, y_i)$, for $i = 1,2,\ldots,n$, are the set of n data values.

To find the coefficients, we need to solve the following system of linear equations.

$$\frac{\partial S_r}{\partial a_j} = 0, \qquad j = 0,1,2,\ldots,m$$

As an example, assume that y is related to the independent variables x_1, x_2 using the following linear form:

(x_1, x_2, y)

$$y = \overbrace{a_0 + a_1 x_1 + a_2 x_2}^{\hat{y}} + e$$

and

$e_1 = y_1 - \hat{y}_1 = y_1 - (a_0 + a_1 x_{1,1} + a_2 x_{2,1})$

$e_2 = y_2 - \hat{y}_2 = y_2 - (a_0 + a_1 x_{1,2} + a_2 x_{2,2})$

$e_3 = y_3 - \hat{y}_3 = y_3 - (a_0 + a_1 x_{1,3} + a_2 x_{2,3})$

$\vdots$

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

Thus,

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_{1,i} (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_{2,i} (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}) = 0$$

Therefore, we have the following system:

$$\begin{bmatrix} n & \sum x_{1,i} & \sum x_{2,i} \\ \sum x_{1,i} & \sum x_{1,i}^2 & \sum x_{1,i} x_{2,i} \\ \sum x_{2,i} & \sum x_{1,i} x_{2,i} & \sum x_{2,i}^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1,i} y_i \\ \sum x_{2,i} y_i \end{bmatrix}$$

Example. The following data were created from the equation $y = 5 + 4x_1 - 3x_2$:

y	x_1	x_2	x_1^2	x_2^2	x_1x_2	x_1y	x_2y
5	0	0					
10	2	1					
9	2.5	2					
0	1	3					
3	4	6					
27	7	2					
54	16.5	14	76.25	54	48	243.5	100

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_{1i} \\ \sum y_i x_{2i} \end{bmatrix}$$

$$\begin{bmatrix} 6 & 16.5 & 14 \\ 16.5 & 76.25 & 48 \\ 14 & 48 & 54 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 54 \\ 243.5 \\ 100 \end{bmatrix} \longrightarrow \begin{cases} a_0 = 5 \\ a_1 = 4 \\ a_2 = -3 \end{cases}$$

$$\hat{y} = 5 + 4x_1 - 3x_2$$

$$\hat{y} = a_0 + a_1 x$$

$$\hat{y} = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_m x^m$$

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

Observation and extension.

We have introduced three types of regression: **simple linear, polynomial, and multiple linear**. In fact, all three belong to the following general linear least-squares model:

$$\hat{y} = a_0 z_0 + a_1 z_1 + \cdots + a_m z_m$$

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

Where z_i's are $m + 1$ basis functions.

$z_0 = \cos x, \quad z_1 = \tan x, \quad z_2 = \ln x$

$\hat{y} = a_0 \cos x + a_1 \tan x + a_2 \ln x$

- Simple regression: $z_0 = 1, z_1 = x$
- Polynomial regression: $z_0 = 1, z_1 = x, z_2 = x^2, \dots, z_m = x^m$

Note that the terminology "linear" refers only to the model's dependence on its parameters. As another example, the z's can be sinusoids, as in

$$y = a_0 + a_1 \cos(\omega x) + a_2 \sin(\omega x)$$

The sum of the squares of the residuals for this model can be defined as

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

where $z_{j,i} = z_j(x_{1,i}, \dots, x_{n,i})$. Again, to find the coefficients, we need to solve the following system of linear equations.

$$\frac{\partial S_r}{\partial a_j} = 0, \qquad j = 0,1,2,\dots,m$$

evaluations of the funct z_0 at the observations.

Let

$$Z = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{m1} \\ z_{02} & z_{12} & \cdots & z_{m2} \\ \vdots & \vdots & & \vdots \\ z_{0n} & z_{1n} & \cdots & z_{mn} \end{bmatrix} \qquad \begin{aligned} a &= \lfloor a_0 \quad a_1 \quad \cdots \quad a_m \rfloor^T \\ y &= \lfloor y_1 \quad y_2 \quad \cdots \quad y_n \rfloor^T \end{aligned}$$

In case $n \geq m + 1$, the solution for coefficient vector will be obtained from solving the following *normal equations*:

$$(Z^T Z) a = Z^T y$$

Example. Use Matlab and $z_0 = 1, z_1 = x, z_2 = x^2$ to fit a quadratic function for the following data.

$$\hat{y} = a_0 z_0 + a_1 z_1 + a_2 z_2$$
$$= a_0 + a_1 x + a_2 x^2$$

x_i	y_i
0	2.1
1	7.7
2	13.6
3	27.2
4	40.9
5	61.1

$z_0 = 1$

$z_1 = x$

$z_2 = x^2$

Solution. Note that the backslash function (that is `w=A\b` for solving $Aw = b$) uses QR factorization which is more robust approach for ill-conditioned problems.

```
>> x = [0 1 2 3 4 5]';

>> y = [2.1 7.7 13.6 27.2 40.9 61.1]';

>> Z = [ones(size(x)) x x.^2]

>> a = (Z'*Z)\(Z'*y)
```

$$Z = \begin{array}{c} \begin{array}{ccc} z_0 & z_1 & z_2 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix} \end{array} \mapsto (Z^T Z)\, a = Z^T y$$

$$\xrightarrow{\text{use MATLAB}} \begin{cases} a_0 = 2.4786 \\ a_1 = 2.3593 \\ a_2 = 1.8607 \end{cases}$$

$$\hat{y} = 2.4784 + 2.3593x + 1.8607x^2$$

References

1. Chapra, Steven C. (2018). *Numerical Methods with* MATLAB *for Engineers and Scientists*, 4th Ed. McGraw Hill.
2. Burden, Richard L., Faires, J. Douglas (2011). *Numerical Analysis*, 9th Ed. Brooks/Cole Cengage Learning