Polynomial Regression Model

University of South Florida

United States of America

kaw@eng.usf.edu

Note: This worksheet demonstrates the use of Maple to illustrate the procedure to regress a given data set to a nonlinear polynomial model.

Introduction

Given n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, least squares method can be used to regress the data to a m^{th} order polynomial.

$y = a_0 + a_1x + a_2x^2 + \ldots + a_mx^m$, m<n
(1)

The residual at each data point is given by

$E_i = y_i - a_0 - a_ix^i \ldots - a_mx_i^m$
(2)

The sum of the square of the residuals is given by

$S_r = \sum E_i^2, \; i = 1..n$
(3)

To find the constants of the polynomial regression model, we put the derivatives with respect to a_i to zero, that is,

$dS_r / da_0 = \sum [2(y_i - a_0 \ldots - a_ix^i \ldots - a_mx_i^m, \;\; i = 1..n)(-1)] = 0$
(4.a)
$dS_r / da_1 = \sum [2(y_i - a_0 \ldots - a_ix^i \ldots - a_mx_i^m, \;\; i = 1..n)(-x_i)] = 0$
(4.b)
. . . .
. . . .

$dS_r / da_m = \sum [2(y_i - a_0 \ldots - a_ix^i \ldots - a_mx_i^m, \;\; i = 1..n)(-x_i^m)] = 0$

(4.m)

Setting those equations in matrix form gives

$$\begin{vmatrix} n & \sum(x_i) & \dots & \Sigma(x_i)^m \\ \Sigma(x_i) & \Sigma(x_i)^2 & \dots & \Sigma(x_i)^{m+1} \\ \dots & \dots & \dots & \dots \\ \Sigma(x_i)^m & \Sigma(x_i)^{m+1} & \dots & \Sigma(x_i)^{2m} \end{vmatrix} \begin{vmatrix} a_0 \\ a_1 \\ \dots \\ a_m \end{vmatrix} = \begin{vmatrix} \Sigma(y_i) \\ \Sigma(x_i y_i) \\ \dots \\ \Sigma(x_i^m y_i) \end{vmatrix}$$

The above simultaneous linear equations are solved for the (m+1) constants $a_0, a_1, \dots, a_m$. To learn more about polynomial regression see the worksheet on Nonlinear Regression.

Section 1: Input data

Below are the input parameters to begin the simulation. This is the only section that requires user input. Once the values are entered, Maple will will generate a polynomial regression model for the given data set. It will also calculate the variance each order of polynomial model specified with the *Low_order* to *High_order* range so that user can determine the optimum order of polynomial model to use.

Input Parameters:
X = array of *x* values
Y = array of *y* values
n = Number of data points
Order_poly = desired order of polynomial regression model
Low_order = Lowest order of polynomial to check for optimum order
High_order = Highest order of polynomial to check for optimum order. ***NOTE: High_order must be less than or equal to (n-1)***.

```
> restart;
  X:=[80,40,0,-40,-80,-120,-160,-200,-240,-280,-320];
  Y:=[6.47,6.24,6,5.72,5.43,5.09,4.72,4.30,3.83,3.33,2.76];
  n:=11;
  Order_poly:=1;
  Low_order:=1;
  High_order:=6;
```

$X := [80, 40, 0, -40, -80, -120, -160, -200, -240, -280, -320]$

$Y := [6.47, 6.24, 6, 5.72, 5.43, 5.09, 4.72, 4.30, 3.83, 3.33, 2.76]$

$n := 11$

$Order\_poly := 1$

$Low\_order := 1$

$High\_order := 6$ **(2.1)**

Section 2: Defining the system of simultaneous linear equations in matrix form

In this section, the coefficient matrix "M" and right hand side vector "C" are calculated and subsequently used to determine the solution vector that contains the coefficients of the polynomial model $a_0, a_1, a_2, \ldots, a_m$.

- **Calculating the coefficient matrix, "M"**

The following procedure defines each value of the coefficient matrix.

```
> #Creating the matrix size according to the order of polynomial model:
  M:=Matrix(1..Order_poly+1,1..Order_poly+1):
  #Determing each value of the first row of matrix "M":
  M[1,1]:=n:
  for i from 2 by 1 to (Order_poly+1) do
       for j from 1 by 1 to n do
            M[1,i]:=M[1,i]+(X[j])^(i-1);
       end do;
  end do;
  #Calculating the remaining values of the coefficent matrix:
  for i from 1 by 1 to Order_poly+1 do
       for k from 2 by 1 to Order_poly+1 do
            for j from 1 by 1 to n do
                 M[k,i]:=M[k,i]+(X[j])^(i+k-2):
            end do:
       end do:
  end do:
  print(`M = `, M);
```

$$M = , \begin{bmatrix} 1.10000 \times 10^{1} & -1.32000 \times 10^{3} \\ -1.32000 \times 10^{3} & 3.34400 \times 10^{5} \end{bmatrix} \quad \textbf{(3.1)}$$

- **Calculating the right hand side vector, "C"**

Below, the right hand side vector "C" is determined.

```
> #Creating the vector size according to the order of polynomial model:
  C:=Matrix(1..Order_poly+1,1):
  #Finding each value of "C":
  for i from 1 by 1 to n do
```

```
        C[1,1]:=C[1,1]+Y[i]:
   end do:
   for i from 2 by 1 to Order_poly+1 do
        for j from 1 by 1 to n do
              C[i,1]:=C[i,1]+(X[j]^(i-1))*Y[j]:
        end do:
   end do:
   print(`C=`,C);
>
```

$$C=, \begin{bmatrix} 53.89 \\ -4856.80 \end{bmatrix} \quad (3.2)$$

Section 3: Solving the system of simultaneous linear equations M*a=C

Now that the right hand side vector "C" and coefficient matrix "M" have been calculated, they can be used to solve for the solution vector "a" which contains the coefficients of the polynomial model as $a = [a_0, a_1, \ldots , a_m]$.

```
> #Using Maple to solve for the system of linear equations:
  with(LinearAlgebra):
  a:=LinearSolve(M,C);
```

$$a := \begin{bmatrix} 5.99682 \\ 9.14773 \times 10^{-3} \end{bmatrix} \quad (4.1)$$

The polynomial regression model is as follows:

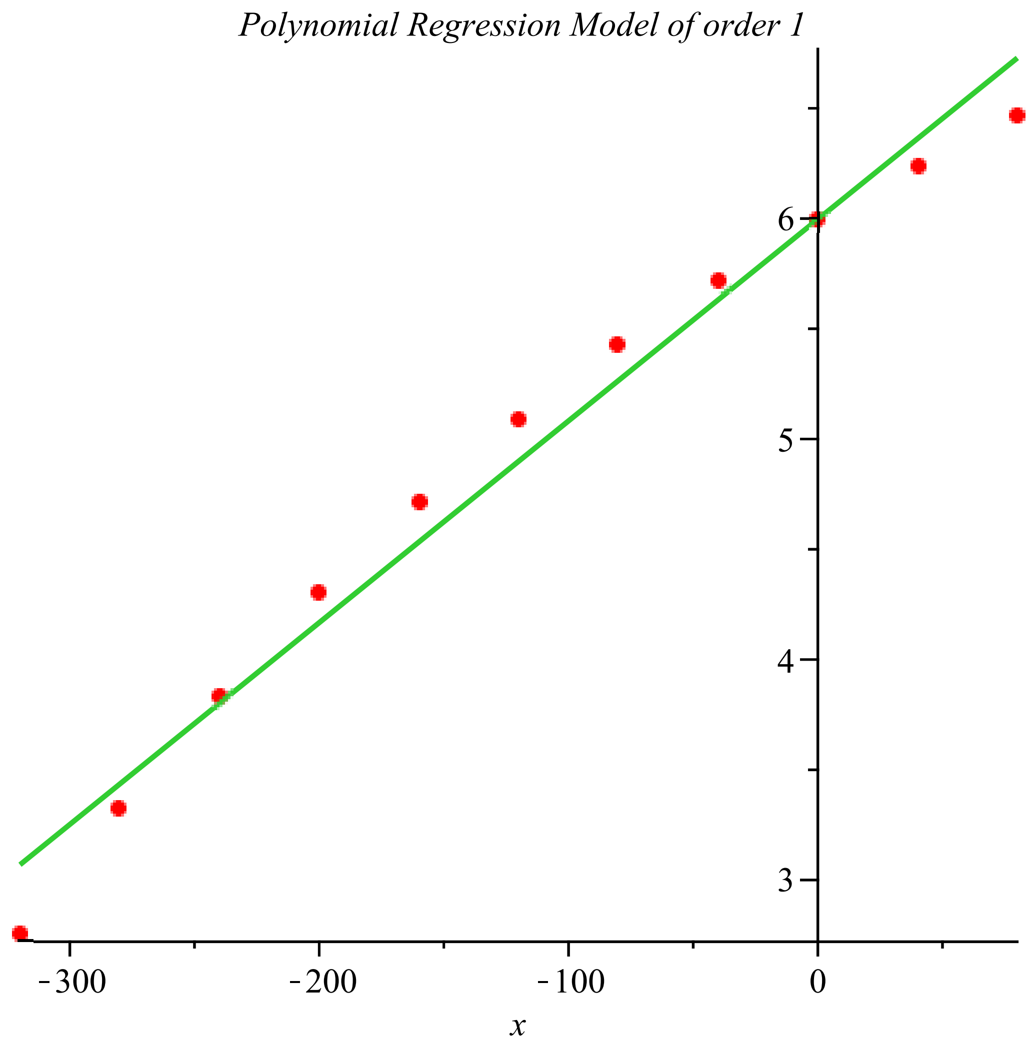
```
> for i from 1 by 1 to n do
       y1:=a[1,1]:
       for j from 1 by 1 to Order_poly do
            y1:=y1+(a[j+1,1]*(x^j)):
       end do:
  end do:
  y:=unapply(y1,x);
```

$$y := x \rightarrow 5.996818181818182 04 + 0.0091477272727272 7266\, x \quad (4.2)$$

The following plot shows the regression model as well as the data points.

```
> observed:=[seq([X[i],Y[i]],i=1..n)];
  predicted:=y(x):
  ttl:=cat(`Polynomial Regression Model of order `,Order_poly):
  plot([observed,predicted],x=X[1]..X[n],title=ttl,style=
  [point,line],symbol=solidcircle,thickness=2,symbolsize=15);
```

observed := [[80, 6.47], [40, 6.24], [0, 6], [-40, 5.72], [-80, 5.43], [-120, 5.09], [-160, 4.72], [-200, 4.30], [-240, 3.83], [-280, 3.33], [-320, 2.76]]



Section 4: Optimum Order

In this section the user can determine the optimum order of the polynomial model by plotting the variance defined as

$$\frac{Sr}{\quad}$$

$n\text{-}(m+1)$

as a function of m, where n is the number of data points, S_r is the sum of the square of residuals and m is the order of the polynomial. The optimum order is considered as to be the one where the value of the variance $S_r/[\,n-(m+1)]$ is minimum or where its value is significantly decreasing.

In the following procedure, an m^{th} order polynomial regression model is calculated for each order specified in the *Low_order* to *High_order* range. Maple then returns the variance of each model. The worksheet does not choose the order of the optimum polynomial for regression for you. Look at the plot of the variance as a function of the order of the polynomial. The optimum polynomial is one after which there is no statistical significant decrease in the variance.

Many a times, the variance may show signs of decreasing and then increasing as a function of the order of the polynomial regression model. Such increases in the variance are normal as the variance is calculated as the ratio between the sum of the squares of the residuals and the difference between the number of data points and number of constants of the polynomial model. Both the numerator and denominator decrease as the order of the polynomial is increased. However, as the order of the polynomial increases, the coefficient matrix in the calculation of the constants of the model becomes more ill-conditioned. This ill-conditioning of the coefficient matrix results in fewer significant digits that can be trusted to be correct in the coefficients of the polynomial model, and hence artificially amplify the value of the variance.

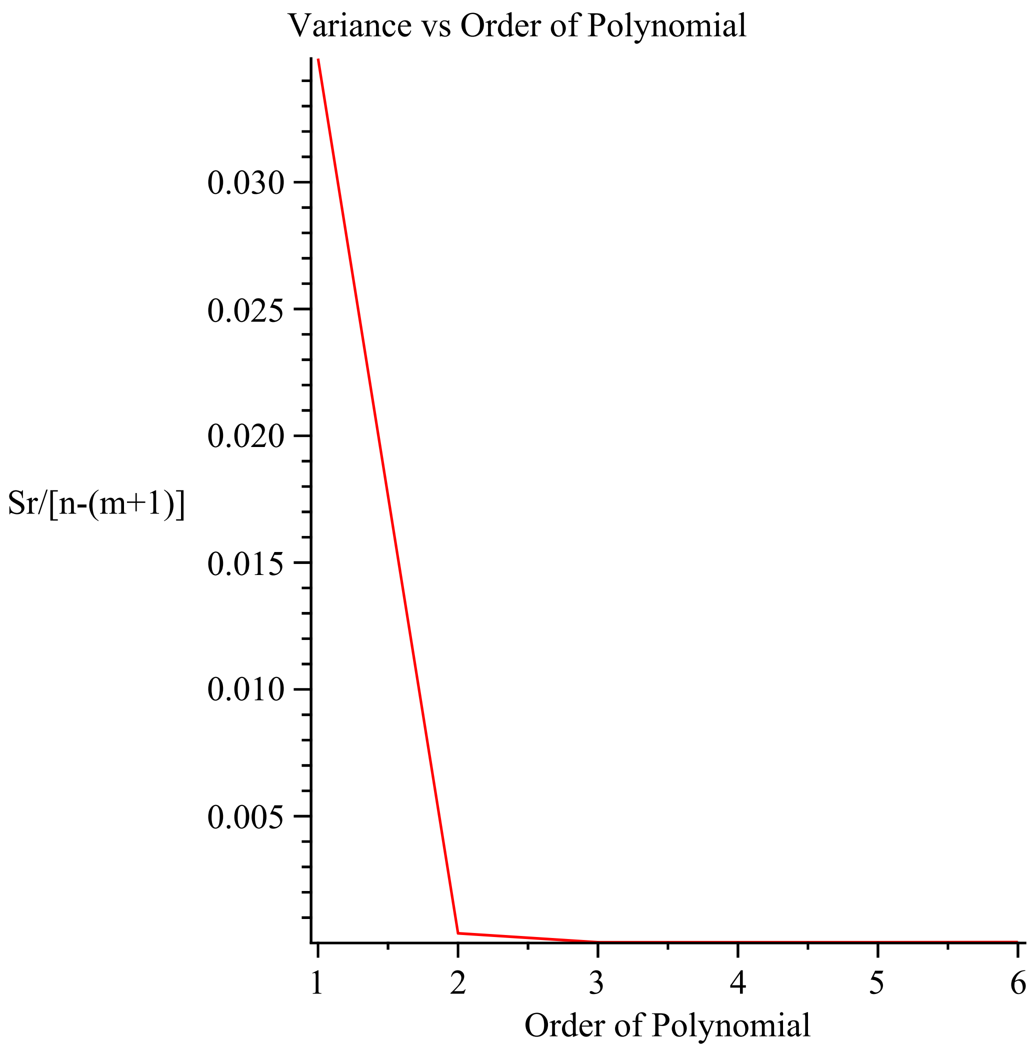
Procedure for calculating variance

```
> Sr=Matrix(1..High_order):
#In the procedure, "m" is the order of polynomial model being calculated, "M" is the coefficient matrix, "c" is the RHS vector, and "A" is the solution vector containing the coefficients of the polynomial model.
for m from Low_order by 1 to High_order do
     M:=Matrix(1..(m+1),1..(m+1)):
     M[1,1]:=n:
     for i from 2 by 1 to m+1 do
          M[1,i]:=0:
          for j from 1 by 1 to n do
                M[1,i]:=M[1,i]+X[j]^(i-1):
          end do:
     end do:
     for i from 1 by 1 to m+1 do
          for k from 2 by 1 to m+1 do
               M[k,i]:=0:
               for j from 1 by 1 to n do
                    M[k,i]:=M[k,i]+X[j]^(i+k-2):
               end do:
          end do:
     end do:
#Calculating the RHS matrix for the given order m:
```

```
        c:=Matrix(1..(m+1),1):
        c[1,1]:=0;
        for i from 1 by 1 to n do
             c[1,1]:=c[1,1]+Y[i]:
        end do:
        for i from 2 by 1 to m+1 do
             c[i,1]:=0:
             for j from 1 by 1 to n do
                  c[i,1]:=c[i,1]+(X[j]^(i-1))*Y[j]:
             end do:
         end do:
```
#Calculating the coefficients of the m^{th} order polynomial model:
```
        A:=LinearSolve(M,c):
```
#Determining *Sr*:
```
        Sr[m]:=0:
        for i from 1 by 1 to n do
             summ:=0;
             for j from 1 by 1 to m do
                  summ:=summ+A[j+1,1]*(X[i]^j);
             end do;
             Sr[m]:=Sr[m]+(Y[i]-(A[1,1]+summ))^2;
        end do;
   end do:
```
#Calculating the variance for the m^{th} order polynomial:
```
   var:=array(1..High_order):
   for i from Low_order by 1 to High_order do
        var[i]:=0;
        var[i]:=Sr[i]/(n-(i+1));
   end do:
   print(`variance = `,var);
   variance:=[seq([i,var[i]],i=Low_order..High_order)]:
   plot(variance,title="Variance vs Order of Polynomial",labels=
   ["Order of Polynomial","Sr/[n-(m+1)]"]);
```

>

variance = , [0.03487222228, 0.0003808857852, 0.00002737262774, 0.00002610722803, 0.00003081585376, 0.00003249486072]



References

[1] *Autar Kaw, Holistic Numerical Methods Institute, http://numericalmethods.eng.usf.edu/mws/gen/06reg, See*
Nonlinear Regression

Conclusion

Using Maple, we are able to regress a given data set to a polynomial model of the m^{th} order.

Question 1: Water is flowing through a pipe of radius 0.5 feet and flow velocity, *v* measurements are made from the center of the wall of the pipe as follows:

Radial Location, *r* (*ft*)	Velocity, *v* (*ft/s*)

0	10
0.08	9.7
0.17	8.9
0.25	7.5
0.32	5.6
0.42	3.1
0.50	0

a) Regress the data to
$v = a_0(1-(r^2/a^2))$, where a is the radius of the pipe.
b) Find the flow rate, Q through the pipe. (Hint: $Q = \int 2\pi r v dr$, $r=0..a$).

Question 2: Thermal expansion coefficient of steel varies with temperature as given in the table below

Temperature, T	Thermal expansion coefficient, α* E-06
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.72
-60	5.58
-80	5.43
-100	5.28
-120	5.09
-140	4.91
-160	4.72
-180	4.52

-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07
-320	2.76
-340	2.45

a) Regress the data to a second order polynomial, $\alpha = a_0 + a_1 T + a_2 T^2$.
b) Find the optimum order of polynomial for the regression model.
c) Find the reduction in the diameter of a steel cylinder of diameter 12.5" if it is cooled from a room temperature of 80°F to -108°F.