
Fitting Orthogonal Polynomial Model

Mohammad Ehsanul Karim <wildscop@yahoo.com>
3rd Year, Course – ASTH 301, Roll – SH 101
Institute of Statistical Research and Training
University of Dhaka, Dhaka- 1000, Bangladesh

We have YEAR values of Z_i ¹. For our convenience of table use, we center them around YEAR 1963. Then we will have the following table-

Y	Z	X = Z-1963
0.52	1964	1
0.60	1965	2
0.68	1966	3
0.73	1967	4
0.75	1968	5
0.79	1969	6
0.73	1970	7
0.85	1971	8
1.04	1972	9
1.49	1973	10
2.89	1974	11
3.41	1975	12
3.30	1976	13
3.01	1977	14
3.16	1978	15
4.33	1979	16
4.42	1980	17
3.00	1981	18
1.77	1982	19

Now, from theory, we know that maximum possible number of order polynomial that may be fit is one less than the number of distinct independent values. Here we do not have any duplicate values of X. Therefore, we have 19 distinct X values and hence maximum possible order of our polynomial is $p = n-1 = 19 - 1 = 18$. Now when the order of the fitted polynomial is the maximum order, then model fits the data exactly so that no residual occurs. So, to have some residual so that we may test our hypotheses about the model, we may consider lowering our model in some orders. Also, from table use consideration, we fit polynomial of 6th order.

Thus, our model for this problem is: $Y = \sum_{j=0}^6 a_j y_j(X_j)$

And, under the Orthogonal Polynomial, the Least Squares estimate of the parameters is given by-

$$\hat{a}_j = \frac{\sum_{i=0}^n Y y_j(X_i)}{\sum_{i=0}^n y_j^2(X_i)}$$

¹ The data was taken from Draper, Smith (1981) exercise (Y) in page 291.

At this point, note that X values are equally spaced and have same number of replicates (r = 1) each time. Therefore, we may use the table of Pearson, Hartley (1958, page 214) for n=19, which is as follows:

y_0	y_1	y_2	y_3	y_4	y_5	y_6
1	-9	51	-204	612	-102	1326
1	-8	34	-68	-68	68	-1768
1	-7	19	28	-388	98	-1222
1	-6	6	89	-453	58	234
1	-5	-5	120	-354	-3	1235
1	-4	-14	126	-168	-54	1352
1	-3	-21	112	42	-79	729
1	-2	-26	83	227	-74	-214
1	-1	-29	44	352	-44	-1012
1	0	-30	0	396	0	-1320
1	1	-29	-44	352	44	-1012
1	2	-26	-83	227	74	-214
1	3	-21	-112	42	79	729
1	4	-14	-126	-168	54	1352
1	5	-5	-120	-354	3	1235
1	6	6	-89	-453	-58	234
1	7	19	-28	-388	-98	-1222
1	8	34	68	-68	-68	-1768
1	9	51	204	612	102	1326
I	1	1	5/6	7/12	1/40	11/120

Then,

$$\begin{aligned}\hat{a}_0 &= \frac{1}{19}(0.52 \times 1 + 0.60 \times 1 + 0.68 \times 1 + 0.73 \times 1 + 0.75 \times 1 + 0.79 \times 1 \\ &+ 0.73 \times 1 + 0.85 \times 1 + 1.04 \times 1 + 1.49 \times 1 + 2.89 \times 1 + 3.41 \times 1 \\ &+ 3.30 \times 1 + 3.01 \times 1 + 3.16 \times 1 + 4.33 \times 1 + 4.42 \times 1 + 3.00 \times 1 + 1.77 \times 1) \\ &= 1.972105 \quad \text{where} \quad \sum_{i=0}^n y_0^2(X_i) = 19\end{aligned}$$

$$\sum_{i=0}^n y_0 y_0(X_i) = 37.47$$

Similarly,

$$\hat{a}_1 = 0.1997193 \quad \text{where} \quad \sum_{i=0}^n y_1^2(X_i) = 570$$

$$\sum_{i=0}^n y_1 y_1(X_i) = 113.84$$

$$\hat{a}_2 = -0.004449359 \quad \text{where} \quad \sum_{i=0}^n y_2^2(X_i) = 13566$$

$$\sum_{i=0}^n y_2 y_2(X_i) = -60.36$$

$$\begin{array}{ll}
\hat{\mathbf{a}}_3 = -0.005429965 \text{ where} & \sum_{i=0}^n \mathbf{y}_3^2(X_i) = 213180 \\
& \sum_{i=0}^n Y \mathbf{y}_3(X_i) = -1157.56 \\
\hat{\mathbf{a}}_4 = -0.0008859454 \text{ where} & \sum_{i=0}^n \mathbf{y}_4^2(X_i) = 2288132 \\
& \sum_{i=0}^n Y \mathbf{y}_4(X_i) = -2027.16 \\
\hat{\mathbf{a}}_5 = -0.0001126217 \text{ where} & \sum_{i=0}^n \mathbf{y}_5^2(X_i) = 89148 \\
& \sum_{i=0}^n Y \mathbf{y}_5(X_i) = -10.04 \\
\hat{\mathbf{a}}_6 = -9.494732 \times 10^{-5} \text{ where} & \sum_{i=0}^n \mathbf{y}_6^2(X_i) = 24515700 \\
& \sum_{i=0}^n Y \mathbf{y}_6(X_i) = -2327.7
\end{array}$$

Now, Total SS = $\sum_{i=1}^n Y_i^2 = 107.7035$

$$SS(\hat{\mathbf{a}}_j) = \frac{\left(\sum_{i=0}^n Y_i \mathbf{y}_j(X_i) \right)^2}{\sum_{i=0}^n \mathbf{y}_j^2(X_i)}$$

Thus,

$$SS(\hat{\mathbf{a}}_0) = \frac{\left(\sum_{i=0}^n Y \mathbf{y}_0(X_i) \right)^2}{\sum_{i=0}^n \mathbf{y}_0^2(X_i)} = \frac{(37.47)^2}{19} = 73.89478$$

Similarly,

$$\begin{array}{ll}
SS(\hat{\mathbf{a}}_1) & = 22.73604 \\
SS(\hat{\mathbf{a}}_2) & = 0.2685633 \\
SS(\hat{\mathbf{a}}_3) & = 6.28551 \\
SS(\hat{\mathbf{a}}_4) & = 1.795953 \\
SS(\hat{\mathbf{a}}_5) & = 0.001130722 \\
SS(\hat{\mathbf{a}}_6) & = 0.2210089
\end{array}$$

Thus, Residual SS = Total SS - $\sum_{j=0}^6 SS(\hat{\mathbf{a}}_j) = 107.7035 - 105.203 = 2.5005$

Estimating Parameters Using SAS

We write the following SAS program to get the estimates:

```
title 'PROC ORTHOREG used with drsm data';
data drsm;
input X Y1 Y2 Y3 Y4;
Year=X-1963;
datalines;
1964 0.52 6.04 0.76 1.99
1965 0.60 7.41 0.90 2.41
1966 0.68 6.24 1.02 2.34
1967 0.73 5.65 0.96 2.34
1968 0.75 6.01 1.03 2.29
1969 0.79 5.94 1.13 2.32
1970 0.73 2.08 1.27 2.52
1971 0.85 6.72 1.38 2.48
1972 1.04 7.51 1.50 2.60
1973 1.49 8.34 1.73 3.15
1974 2.89 3.27 1.93 2.14
1975 3.41 4.32 2.00 3.33
1976 3.30 10.08 2.28 3.62
1977 3.01 11.62 2.50 3.03
1978 3.16 12.24 2.93 4.34
1979 4.33 10.04 3.26 6.02
1980 4.42 -2.65 3.48 5.30
1981 3.00 1.07 3.63 2.64
1982 1.77 3.09 4.26 2.67
;
proc iml;
  use drsm;
  read all var {X} into year;
  read all var {Y1} into Y1;
/*return orthogonal polynomial design matrix of order6 */
  x = orpol(year-1963,6);
/* Labels to use for design matrix */
  order = {"Intercept", "1st order", "2nd order", "3rd
order", "4th order", "5th order", "6th order"};
  ds_ssqr={19,570,13566,213180,2288132,89148,24515700};
  x_new = fuzz(x`#sqrt(ds_ssqr))`;
/* Fit regression */
  run regress(x_new,Y1,order,,,,);
quit;
```

Note that, the code provides estimates of unnormalized orthogonal polynomials.

Replacing -

```
ds_ssqr = {19,570,13566,213180,2288132,89148,24515700};
x_new = fuzz(x`#sqrt(ds_ssqr))`;
run regress(x_new,Y1,order,,,,);
```

by-

```
run regress(x,Y1,order,,,,);
```

will provide estimates of normalized Orthonormal Polynomial. This can also be found using *S-plus* or *R* command : `orth.fit<-lm(Y~poly(X,degree =6))2`.

However, the output of the unnormalized orthogonal polynomial codes will be as follows:

```

1                                PROC ORTHOREG used with drsm data
                                00:35

NAME          B          STDB          T          PROBT
Intercept  1.9721053  0.104724 18.831448 2.814E-10
1st order  0.1997193 0.0191199 10.445621 2.2315E-7
2nd order  -0.004449 0.0039192 -1.135273 0.2784283
3rd order  -0.00543 0.0009887 -5.49221 0.000138
4th order  -0.000886 0.0003018 -2.935785 0.0124692
5th order  -0.000113 0.0015289 -0.073664 0.9424915
6th order  -0.000095 0.0000922 -1.029868 0.3233833

                                Covariance of Estimates

order  COVB      Intercept 1st order 2nd order 3rd order 4th order 5th order 6th
0      Intercept      0.011          0          0          0          0          0
0      1st order          0      0.0004          0          0          0          0
0      2nd order          0          0      154E-7          0          0          0
0      3rd order          0          0          0      977E-9          0          0
0      4th order          0          0          0          0      91E-9          0
0      5th order          0          0          0          0          0      234E-8
0      6th order          0          0          0          0          0          0      85E-
10

                                Correlation of Estimates

order  CORRB      Intercept 1st order 2nd order 3rd order 4th order 5th order 6th
0      Intercept          1          0          0          0          0          0
0      1st order          0          1          0          0          0          0
0      2nd order          0          0          1          0          0          0
0      3rd order          0          0          0          1          0          0
0      4th order          0          0          0          0          1          0
0      5th order          0          0          0          0          0          1
0      6th order          0          0          0          0          0          0
1

```

² Special thanks to Douglas Bates, Dale McLerran, Prof Brian Ripley, Nick Cox for making a beginner like me to understand such details.

ANOVA table

Source of Variation	df	SS	MS	F _{cal}	F _{tab} (5%)
$\hat{\mathbf{a}}_0$	1	73.89478	73.8947842105263	354.6234	4.7472
$\hat{\mathbf{a}}_1$	1	22.73604	22.7360449122807	109.111	4.7472
$\hat{\mathbf{a}}_2$	1	0.2685633	0.268563290579389	1.288844	4.7472*
$\hat{\mathbf{a}}_3$	1	6.28551	6.28551061825687	30.16437	4.7472
$\hat{\mathbf{a}}_4$	1	1.795953	1.79595305935147	8.618836	4.7472
$\hat{\mathbf{a}}_5$	1	0.001130722	0.001130721945528	0.005426371	4.7472*
$\hat{\mathbf{a}}_6$	1	0.2210089	0.221008876	1.060629	4.7472*
Residual	12	2.500504	0.2083754		

Here, $R^2 = (\text{Reg SS}) / (\text{Total SS}) = 105.203 / 107.7035 = 0.9767834$ and
Residual df = $n - k - 1 = 19 - 6 - 1 = 12$

As we observed that the terms of the fourth and lower order (other than 2nd) account for most of the variation in the data, we can adopt the model as-

$$\begin{aligned}\hat{Y} &= \hat{\mathbf{a}}_0 + \hat{\mathbf{a}}_1 \mathbf{y}_1(X_i) + \hat{\mathbf{a}}_3 \mathbf{y}_3(X_i) + \hat{\mathbf{a}}_4 \mathbf{y}_4(X_i) \\ &= \begin{bmatrix} 1.972105 \\ +0.1997193 \mathbf{y}_1(X_i) \\ -0.005429965 \mathbf{y}_3(X_i) \\ -0.0008859454 \mathbf{y}_4(X_i) \end{bmatrix}\end{aligned}$$

We can rearrange the ANOVA table as-

New ANOVA table

Source of Variation	df	SS	MS	F _{cal}	F _{tab} (5%)
$\hat{\mathbf{a}}_0$	1	73.89478	73.8947842105263	370.560008	4.5431
$\hat{\mathbf{a}}_1$	1	22.73604	22.7360449122807	114.014393	4.5431
$\hat{\mathbf{a}}_3$	1	6.28551	6.28551061825687	31.519936	4.5431
$\hat{\mathbf{a}}_4$	1	1.795953	1.79595305935147	9.006162	4.5431
Residual	15	2.991207	0.1994138		

Here Residual SS = Total SS - $\sum_{j=0}^4 SS(\hat{\mathbf{a}}_j) = 107.7035 - 104.7123 =$
 2.991207 and Residual df = $n - k - 1 = 19 - 3 - 1 = 15$.

Therefore, the above model is final and retained since no parameters are rejected at 5% significance level. Here $R^2 = (\text{Reg SS}) / (\text{Total SS}) = 0.9722274$ is a bit less than the previous one, but for sake of simplicity of the model and interpretation, we choose this model.

Now, in order to obtain the fitted equation in terms of the original variables, we have to first substitute for the y_j and relate these to X_i 's. The formula for the Orthogonal Polynomial upto fourth order for equally spaced X 's for $n = 19$ is-

$$\begin{aligned}
 y_0(X_i) &= 1 \\
 y_1(X_i) &= I_1 X = X \\
 y_2(X_i) &= I_2 \left[X^2 - \left(\frac{n^2 - 1}{12} \right) \right] = \left[X^2 - \left(\frac{19^2 - 1}{12} \right) \right] = X^2 - 30 \\
 y_3(X_i) &= I_3 \left[X^3 - \left(\frac{3n^2 - 7}{20} \right) X \right] = \frac{5}{6} \left[X^3 - \left(\frac{19^2 - 7}{20} \right) X \right] = \frac{5}{6} X^3 - \frac{177}{12} X \\
 y_4(X_i) &= \frac{7}{12} \left[X^4 - \frac{1}{14} (19^2 - 13) X^2 + \frac{3}{560} (19^2 - 1)(19^2 - 9) \right] \\
 &= \frac{7}{12} X^4 - \frac{609}{42} X^2 + 396
 \end{aligned}$$

Therefore, Z and X corresponds as follows –

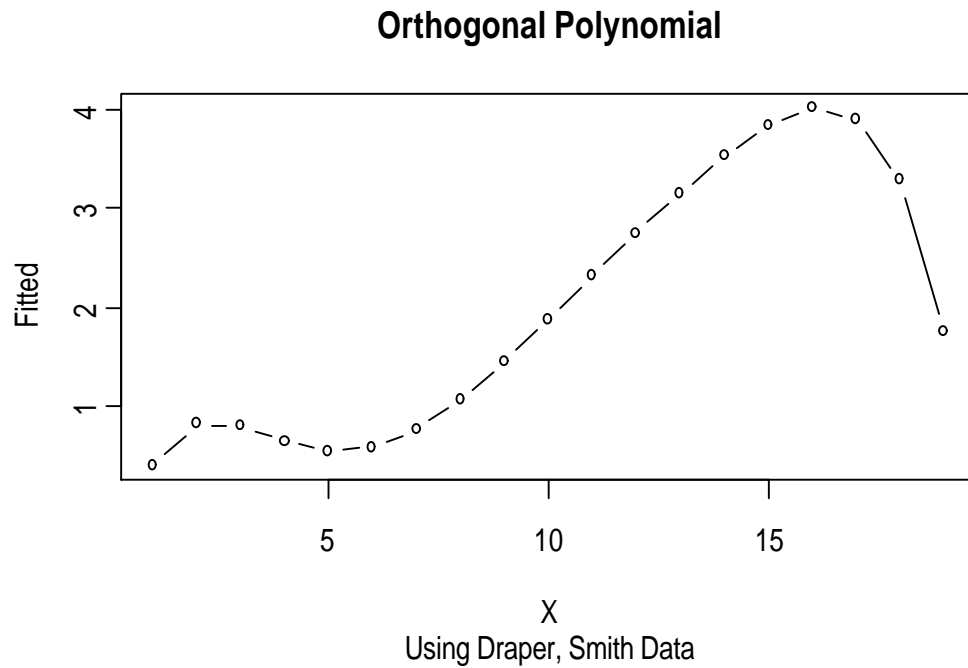
$$\begin{aligned}
 y_1(X_i) = X_i: & -9 \ -8 \ -7 \ -6 \ -5 \ -4 \ -3 \ -2 \ -1 \ 0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \\
 = Z_i: & 1964 \ 1965 \ 1966 \ 1967 \ 1968 \ 1969 \ 1970 \ 1971 \ 1972 \ 1973 \ 1974 \ 1975 \ 1976 \ 1977 \ 1978 \ 1979 \ 1980 \ 1981 \ 1982
 \end{aligned}$$

where the coding is found to be $X = Z - 1973$ since there X is 0.

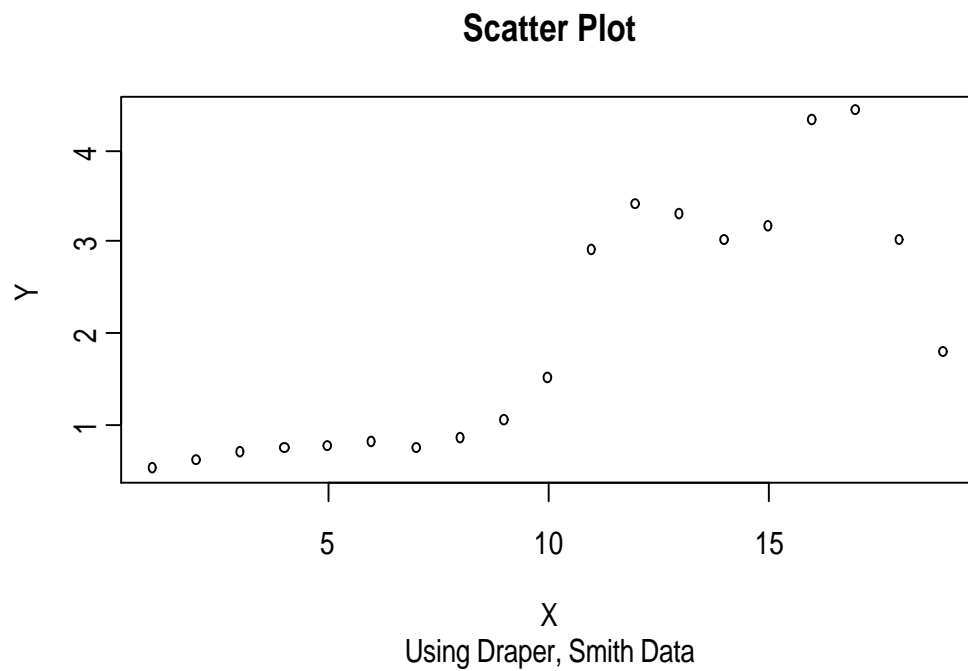
Thus the fitted polynomial is –

$$\begin{aligned}
 \hat{Y} &= \hat{a}_0 + \hat{a}_1 y_1(X_i) + \hat{a}_3 y_3(X_i) + \hat{a}_4 y_4(X_i) \\
 &= \left\{ \begin{array}{l} 1.972105 \\ +0.1997193 y_1(X_i) \\ -0.005429965 y_3(X_i) \\ -0.0008859454 y_4(X_i) \end{array} \right\} \\
 &= 1.972105 + 0.1997193X - 0.005429965 \left(\frac{5}{6} X^3 - \frac{177}{12} X \right) - \\
 &\quad 0.0008859454 \left(\frac{7}{12} X^4 - \frac{609}{42} X^2 + 396 \right)
 \end{aligned}$$

And, our Orthogonal Polynomial Model looks like the following graph (after normalizing³)-



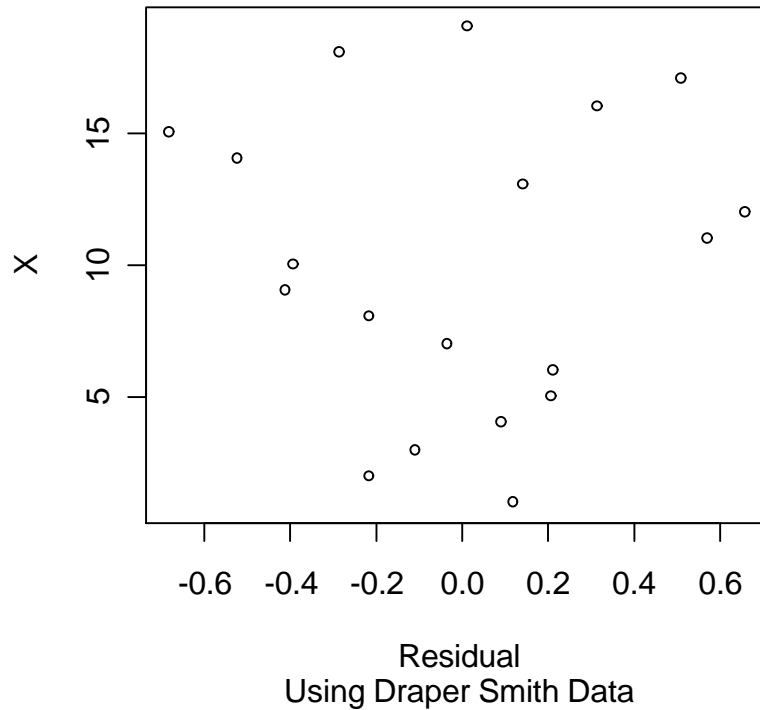
Also, for comparison purposes, we also show the original scatter plot -



³ Dividing each parameters (except the intercept term) by respective Sum of Squares obtained from table.

Checking the following Residual plots, we find them more or less random, and conclude that this model that we have fitted is a good one (not only based on R^2 value, but also on validity grounds).

Residual Vs. X



Residual Vs. Fitted Y

