CHAPTER 22

Orthogonal Polynomials and Summary Data

22.1. INTRODUCTION

In this chapter we bring together two topics with specialized usages in regression situations.

1. Orthogonal polynomials were much used in precomputer days to fit polynomial models of any order to a single predictor X, mostly to equally spaced predictor values. This was achieved by replacing the columns of the $\mathbf{X}$ matrix that would arise from the predictors $(1, X, X^2, X^3, \ldots)$ by alternative columns that were orthogonal and generated by special polynomials of (zero, first, second, third, . . .) order in X. The maximum possible order is, of course, $n - 1$ if n observations are available. Such polynomials could be worked out once and for all. The best-known published tables are by Fisher and Yates (1964) and Pearson and Hartley (1958). Once the orthogonal polynomial fit has been made, the original polynomial form can then be reconstructed, if desired.

Even in the computer age, orthogonal polynomials could be useful, particularly with computer programs that have troublesome round-off errors. These polynomials are also useful more generally, however. They can be used in any situation where orthogonal columns with integer values are needed, for example, to span an estimation space or an error space, to split a sum of squares up into orthogonal components, or to obtain a set of orthogonal dummy variable columns.

Orthogonal polynomials are discussed in Section 22.2.

2. Summary data are data in which the Y's from individual repeat runs are not available because the data have been reduced to sample averages and sample variance estimates in writing a paper or report. It is possible to carry out a large portion of the regression calculations nevertheless. This is discussed in Section 22.3.

22.2. ORTHOGONAL POLYNOMIALS

Orthogonal polynomials are used to fit a polynomial model of any order in one variable. The idea is as follows. Suppose we have n observations (X_i, Y_i), $i = 1, 2, \ldots, n$, where X is a predictor variable and Y is a response variable, and we wish to fit the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \epsilon. \qquad (22.2.1)$$

In general, the columns of the resulting **X** matrix will not be orthogonal. If later we wish to add another term $\beta_{p+1}X^{p+1}$, changes will usually occur in the estimates of all the other coefficients. However, if we can construct polynomials of the form

$$\begin{array}{ll}
\psi_0(X_i) = 1 & \text{zero-order polynomial} \\
\psi_1(X_i) = P_1X_i + Q_1 & \text{first-order} \\
\psi_2(X_i) = P_2X_i^2 + Q_2X_i + R_2 & \text{second-order} \\
\vdots & \\
\psi_r(X_i) = P_rX_i^r + Q_rX_i^{r-1} + \cdots + T_r & r\text{th order} \\
\vdots &
\end{array}$$

with the property that they are *orthogonal polynomials*, that is,

$$\sum_{i=1}^{n} \psi_j(X_i)\psi_l(X_i) = 0 \qquad (j \neq l), \tag{22.2.2}$$

for all j, $l < n - 1$, we can rewrite the model as

$$Y = \alpha_0\psi_0(X) + \alpha_1\psi_1(X) + \cdots + \alpha_p\psi_p(X) + \epsilon. \tag{22.2.3}$$

In this case we shall have

$$\mathbf{X} = \begin{bmatrix} 1 & \psi_1(X_1) & \psi_2(X_1) & \cdots & \psi_p(X_1) \\ 1 & \psi_1(X_2) & \psi_2(X_2) & \cdots & \psi_p(X_2) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \psi_1(X_n) & \psi_2(X_n) & \cdots & \psi_p(X_n) \end{bmatrix} \tag{22.2.4}$$

so that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} A_{00} & & \mathbf{0} & \\ & A_{11} & & \\ & & A_{22} & \\ \mathbf{0} & & & \ddots & \\ & & & & A_{pp} \end{bmatrix}, \tag{22.2.5}$$

where $A_{jj} = \Sigma_{i=1}^{n}\{\psi_j(X_i)\}^2$ since all off-diagonal terms vanish, by Eq. (22.2.2). Since the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is also diagonal and is obtained by inverting each element separately, the least squares procedure provides, as an estimate of α_j, the quantity

$$\begin{aligned} a_j &= \frac{\Sigma_{i=1}^{n} Y_i\psi_j(X_i)}{\Sigma_{i=1}^{n} [\psi_j(X_i)]^2}, \qquad j = 0, 1, 2, \ldots, p, \\ &= \frac{A_{jY}}{A_{jj}} \end{aligned} \tag{22.2.6}$$

with an obvious notation. Since $V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ for the general regression model, it follows that the variance of a_j is

$$V(a_j) = \frac{\sigma^2}{A_{jj}}, \tag{22.2.7}$$

where σ^2 is usually estimated from the analysis of variance table. To obtain the entries in the table we evaluate the sum of squares due to a_j from

$$\text{SS}(a_j) = \frac{A_{jY}^2}{A_{jj}} \tag{22.2.8}$$

and so obtain an analysis of variance table as in Table 22.1. If the model is correct, s^2 estimates σ^2. Usually we pool, with the residual, sums of squares whose mean squares are not significantly larger than s^2 to obtain an estimate of σ^2 based on more degrees of freedom. Note that if it were desired to add another term $\alpha_{p+1}\psi_{p+1}(X)$ to Eq. (22.2.3) no recomputation would be necessary for the coefficients already obtained, due to the orthogonality of the polynomials. Thus higher and higher order polynomials can be fitted with ease and the process terminated when a suitably fitting equation is found.

The $\psi_j(X_i)$ can be constructed and the procedure above can be carried out for any values of the X_i. However, when the X_i are *not* equally spaced, polynomials must be specially constructed. See, for example, Wishart and Metakides (1953) and Robson (1959). If the X_i are equally spaced, published tables can be employed. The actual numerical values of $\psi_j(X_j)$ and A_{jj} as well as the general functional forms of $\psi_j(X)$, for $j = 1, 2, \ldots, 6$ and for $n \leq 52$, are given in Pearson and Hartley (1958). This means that provided $p \leq 6$ we have the elements of the $\mathbf{X}$ and $\mathbf{X}'\mathbf{X}$ matrices. Similar values for $j \leq 5$ and $n \leq 75$ appear in Fisher and Yates (1964). When $p > 6$, more extensive tables such as De Lury (1960) are needed. See also Wilkie (1965). A short table of orthogonal polynomials for equally spaced data for values of n as high as 12 is given later in this section.

Although orthogonal polynomials are often recommended only when pocket calculators are used, Bright and Dawkins (1965) concluded that even when a computer is available, and especially when the polynomial is of high order, orthogonal polynomials are worthwhile. Using them provides greater computing accuracy and reduced computing times and can avoid rejection of columns of powers of X that are highly correlated with other columns of powers of X.

We illustrate the application of orthogonal polynomials by an example.

Example

The Archer Daniels Midland Company reported net income per share for the years 1986–1993 as follows:

TABLE 22.1. Analysis of Variance Table

Source	df	SS	MS
a_0 (mean)	1	SS(a_0)	—
a_1	1	SS(a_1)	SS(a_1)
a_2	1	SS(a_2)	SS(a_2)
⋮	⋮	⋮	⋮
a_p	1	SS(a_p)	SS(a_p)
Residual	$n - p - 1$	By subtraction	s^2
Total	n	$\sum_{i=1}^{n} Y_i^2$	

Year (Z_i)	1986	1987	1988	1989	1990	1991	1992	1993
Income per share (\$$Y_i$)	0.70	0.77	1.02	1.24	1.41	1.35	1.47	1.66

Fit a polynomial of suitable order that will provide a satisfactory approximating function for these data.

Solution. We ignore the year values Z_i for the moment except to note that they are equally spaced. From a table of orthogonal polynomials we find values $\psi_j(X_i)$ corresponding to $n = 8$ observations as shown in Table 22.2.

Note that $A_{00} = n = 8$. We consider the model

$$Y = \sum_{j=0}^{6} \alpha_j \psi_j(X).$$

Using Eq. (22.2.6) we obtain

$$a_0 = 1.2025, \qquad a_1 = 0.067738, \qquad a_2 = -0.009524,$$

$$a_3 = 0.001515, \qquad a_4 = 0.006721, \qquad a_5 = -0.000559,$$

$$a_6 = -0.002879.$$

The fitted model (with $R^2 = 0.997$) is thus

$$\hat{Y} = \sum_{j=0}^{6} a_j \psi_j(X),$$

where the a_j are as shown and the $\psi_j(X)$ are found in tables. (We shall do this in a moment, after reducing the equation.) The analysis of variance table is shown in Table 22.3. All the sums of squares for $a_1, a_2, \ldots, a_6$ are orthogonal and their values do not depend on the order of entry of the polynomial terms.

Note. When the order of the fitted polynomial is $p = n - 1$, the maximum order, the model fits the data exactly and no residual occurs. Here $p = n - 2$ and so the residual is SS(a_7) in fact.

If an external estimate of σ^2 were available we could compare all the mean squares with it. Alternatively, we note that the fourth-, fifth-, and sixth-order sums of squares are very small and the corresponding terms can be dropped immediately. It is mechanically tempting to declare the fourth-order term significant by performing the F-test with (1, 3) degrees of freedom $F = \{0.028/1\}\{(0.001 + 0.002 + 0.003)/3\} = 14$, which exceeds $F(1, 3, 0.95) = 10.13$. Fitting a fourth-order polynomial to eight data points

TABLE 22.2. The Calculation Table for the Example

Y	ψ_0	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
0.70	1	−7	7	−7	7	−7	1
0.77	1	−5	1	5	−13	23	−5
1.02	1	−3	−3	7	−3	−17	9
1.24	1	−1	−5	3	9	−15	−5
1.41	1	1	−5	−3	9	15	−5
1.35	1	3	−3	−7	−3	17	9
1.47	1	5	1	−5	−13	−23	−5
1.66	1	7	7	7	7	7	1
A_{jj}	8	168	168	264	616	2184	264

TABLE 22.3. ANOVA Table for the Example

Source	df	SS	MS
a_1	1	0.771	(same as SS)
a_2	1	0.015	
a_3	1	0.001	
a_4	1	0.028	
a_5	1	0.001	
a_6	1	0.002	
Residual	1	0.003	
Total, corrected	7	0.820	

is not very sensible in general, however, and once we get past this point, it seems sensible to reduce further. The straight line model $Y = 1.2025 + 0.0067738\psi_1(X)$ explains $R^2 = 0.940$ of the variation about the mean and the residual mean square from this model is $s^2 = 0.00818$. A plot of the data and fitted line shows this model to be a reasonable fit.

In order to obtain the fitted equation in terms of the original variables we have first to substitute for the ψ_j and relate these to the Z's. From a table of orthogonal polynomials with $n = 8$ (e.g., the information provided later in this section),

$$\psi_0(X) = 1, \qquad \psi_1(X) = 2X, \qquad \psi_2(X) = X^2 - \frac{21}{4},$$

$$\psi_3(X) = \frac{2}{3}\left[X^3 - \frac{37}{4}X\right], \qquad \text{and so on.}$$

Z and X correspond as follows:

$\psi_1(X_i) = 2X_i$:	-7	-5	-3	-1	1	3	5	7
X_i:	$-\frac{7}{2}$	$-\frac{5}{2}$	$-\frac{3}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$
Z_i:	1986	1987	1988	1989	1990	1991	1992	1993

Clearly the required coding is

$$X = Z - 1989\tfrac{1}{2}.$$

The fitted polynomial is thus

$$\hat{Y} = 1.2025 + 0.0067738(2)(Z - 1989\tfrac{1}{2})$$

or

$$\hat{Y} = 1.2025 + 0.0135476(Z - 1989\tfrac{1}{2}).$$

This can be rearranged as a straight line fit in Z but the above form is as convenient to substitute into to obtain fitted values and residuals, in this case. It is recommended that the reader plot the data and fitted values and also examine the residuals, to complete the analysis.

Orthogonal Polynomials for $n = 3, \ldots, 12$

The formulas for the orthogonal polynomials $\psi_j(X)$ up to order six for equally spaced X's and for any value of n are as follows:

$$\psi_0(X) = 1,$$

$$\psi_1(X) = \lambda_1 X,$$

$$\psi_2(X) = \lambda_2\{X^2 - \tfrac{1}{12}(n^2 - 1)\},$$

$$\psi_3(X) = \lambda_3\{X^3 - \tfrac{1}{20}(3n^2 - 7)X\},$$

$$\psi_4(X) = \lambda_4\{X^4 - \tfrac{1}{14}(3n^2 - 13)X^2 + \tfrac{3}{560}(n^2 - 1)(n^2 - 9)\},$$

$$\psi_5(X) = \lambda_5\{X^5 - \tfrac{5}{18}(n^2 - 7)X^3 + \tfrac{1}{1008}(15n^4 - 230n^2 + 407)X\},$$

$$\psi_6(X) = \lambda_6\{X^6 - \tfrac{5}{44}(3n^2 - 31)X^4 + \tfrac{1}{176}(5n^4 - 110n^2 + 329)X^2 - \tfrac{5}{14784}(n^2 - 1)(n^2 - 9)(n^2 - 25)\}.$$

(If $n \leq 6$, we go only as far as $j = n - 1$.) The appropriate λ-values are given at the bottom of each column and are chosen to ensure that the values in the tables are all integers. Above the λ's are the sums of squares of the column entries, that is, the A_{jj} of Eq. (22.2.5). Note that although X does not appear as such in the tables, it is always true that $\psi_1 = \lambda_1 X$, where λ_1 is 1 for odd n and 2 for even n, so we can read off X as ψ_1/λ_1, when needed. All ψ_j columns are symmetric for even j, antisymmetric for odd j.

For a Fortran subroutine that generates orthogonal polynomials, see Cooper (1971).]

Table of Orthogonal Polynomial Coefficients

$n = 3$		$n = 4$			$n = 5$				$n = 6$				
ψ_1	ψ_2	ψ_1	ψ_2	ψ_3	ψ_1	ψ_2	ψ_3	ψ_4	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5
−1	1	−3	1	−1	−2	2	−1	1	−5	5	−5	1	−1
0	−2	−1	−1	3	−1	−1	2	−4	−3	−1	7	−3	5
1	1	1	−1	−3	0	−2	0	6	−1	−4	4	2	−10
		3	1	1	1	−1	−2	−4	1	−4	−4	2	10
					2	2	1	1	3	−1	−7	−3	−5
									5	5	5	1	1
2	6	20	4	20	10	14	10	70	70	84	180	28	252
1	3	2	1	$\frac{10}{3}$	1	1	$\frac{5}{6}$	$\frac{35}{12}$	2	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{7}{12}$	$\frac{21}{10}$

$n = 7$						$n = 8$					
ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
−3	5	−1	3	−1	1	−7	7	−7	7	−7	1
−2	0	1	−7	4	−6	−5	1	5	−13	23	−5
−1	−3	1	1	−5	15	−3	−3	7	−3	−17	9
0	−4	0	6	0	−20	−1	−5	3	9	−15	−5
1	−3	−1	1	5	15	1	−5	−3	9	15	−5
2	0	−1	−7	−4	−6	3	−3	−7	−3	17	9
3	5	1	3	1	1	5	1	−5	−13	−23	−5
						7	7	7	7	7	1
28	84	6	154	84	924	168	168	264	616	2184	264
1	1	$\frac{1}{6}$	$\frac{7}{12}$	$\frac{7}{20}$	$\frac{77}{60}$	2	1	$\frac{2}{3}$	$\frac{7}{12}$	$\frac{7}{10}$	$\frac{11}{60}$

Table of Orthogonal Polynomial Coefficients (*Continued*)

$n = 9$						$n = 10$					
ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
−4	28	−14	14	−4	4	−9	6	−42	18	−6	3
−3	7	7	−21	11	−17	−7	2	14	−22	14	−11
−2	−8	13	−11	−4	22	−5	−1	35	−17	−1	10
−1	−17	9	9	−9	1	−3	−3	31	3	−11	6
0	−20	0	18	0	−20	−1	−4	12	18	−6	−8
1	−17	−9	9	9	1	1	−4	−12	18	6	−8
2	−8	−13	−11	4	22	3	−3	−31	3	11	6
3	7	−7	−21	−11	−17	5	−1	−35	−17	1	10
4	28	14	14	4	4	7	2	−14	−22	−14	−11
						9	6	42	18	6	3
60	2772	990	2002	468	1980	330	132	8580	2860	780	660
1	3	$\frac{5}{6}$	$\frac{7}{12}$	$\frac{3}{20}$	$\frac{11}{60}$	2	$\frac{1}{2}$	$\frac{5}{3}$	$\frac{5}{12}$	$\frac{1}{10}$	$\frac{11}{240}$

$n = 11$						$n = 12$					
ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
−5	15	−30	6	−3	15	−11	55	−33	33	−33	11
−4	6	6	−6	6	−48	−9	25	3	−27	57	−31
−3	−1	22	−6	1	29	−7	1	21	−33	21	11
−2	−6	23	−1	−4	36	−5	−17	25	−13	−29	25
−1	−9	14	4	−4	−12	−3	−29	19	12	−44	4
0	−10	0	6	0	−40	−1	−35	7	28	−20	−20
1	−9	−14	4	4	−12	1	−35	−7	28	20	−20
2	−6	−23	−1	4	36	3	−29	−19	12	44	4
3	−1	−22	−6	−1	29	5	−17	−25	−13	29	25
4	6	−6	−6	−6	−48	7	1	−21	−33	−21	11
5	15	30	6	3	15	9	25	−3	−27	−57	−31
						11	55	33	33	33	11
110	858	4290	286	156	11220	572	12012	5148	8008	15912	4488
1	1	$\frac{5}{6}$	$\frac{1}{12}$	$\frac{1}{40}$	$\frac{11}{120}$	2	3	$\frac{2}{3}$	$\frac{7}{24}$	$\frac{3}{20}$	$\frac{11}{360}$

22.3. REGRESSION ANALYSIS OF SUMMARY DATA

Suppose we have k sets of repeat observations $\{Y_{iu}, u = 1, 2, \ldots, n_i\}$, $i = 1, 2, \ldots, k$, but, because of condensation of the data, the individual repeats are not shown and only the k averages $\overline{Y}_i$ and the k sample variance estimates (of σ^2)

$$s_i^2 = \sum_{u=1}^{n_i} (Y_{iu} - \overline{Y}_i)^2/(n_i - 1)$$

are given. How do we proceed? *For purposes of obtaining the regression coefficients, we can simply work as if* $Y_{iu} = \overline{Y}_i$, that is, as if every set of repeat observations consisted of an equal number of observations all equal to the average value. This is clear if we think of what happens to the ith set of repeats in the matrix product $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. For the $\mathbf{X}'\mathbf{Y}$ portion we have

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \ldots & \ldots & 1 & 1 & \ldots & 1 & \ldots & \ldots \\ \ldots & \ldots & a & a & \ldots & a & \ldots & \ldots \\ \ldots & \ldots & b & b & \ldots & b & \ldots & \ldots \\ & & \vdots & & & & & \\ \ldots & \ldots & z & z & \ldots & z & \ldots & \ldots \end{bmatrix} \begin{bmatrix} \ldots \\ \\ \\ \\ \ldots \\ Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \\ \vdots \\ \ldots \end{bmatrix}.$$

All the a's, b's, and so on, are equal because of the repeat runs. Thus all the contributions of the ith set of repeat observations to the $\mathbf{X}'\mathbf{Y}$ product take forms like,

$$\begin{aligned} aY_{i1} + aY_{i2} + \cdots + aY_{in_i} &= a\sum_{u=1}^{n_i} Y_{iu} \\ &= an_i\,\overline{Y}_i \\ &= a\overline{Y}_i + a\overline{Y}_i + \cdots + a\overline{Y}_i \end{aligned}$$

so that replacement of all Y_{iu} by $\overline{Y}_i$ will not alter the calculation for the vector $\mathbf{b}$.

Use of the $\overline{Y}_i$ instead of the individual Y_{iu} will, however, lead to the wrong total corrected sum of squares. For example, if $\overline{Y}$ is the overall mean of the observations, the contribution to the (wrong) corrected sum of squares from the ith set of observations will be

$$\sum_{u=1}^{n_i} (\overline{Y}_i - \overline{Y})^2 = n_i(\overline{Y}_i - \overline{Y})^2,$$

whereas we actually need

$$\sum_{u=1}^{n_i} (Y_{iu} - \overline{Y})^2.$$

However, we can easily show that

$$\begin{aligned} \sum_{u=1}^{n_i} (Y_{iu} - \overline{Y})^2 &= \sum_{u=1}^{n_i} \{(Y_{iu} - \overline{Y}_i) + (\overline{Y}_i - \overline{Y})\}^2 \\ &= \sum_{u=1}^{n_i} (Y_{iu} - \overline{Y}_i)^2 + n_i(\overline{Y}_i - \overline{Y})^2, \end{aligned}$$

the cross-product term canceling on applying the summation. Thus to obtain the *correct* contribution to the corrected sum of squares for the ith set of repeats, we need to add to $n_i(\overline{Y}_i - \overline{Y})^2$ the quantity

$$\sum_{u=1}^{n_i} (Y_{iu} - \overline{Y}_i)^2 = (n_i - 1)s_i^2.$$

(In some data, only the s_i are given, and it must be remembered that these must be squared before multiplying by $n_i - 1$.)

The quantities $(n_i - 1)s_i^2$, $i = 1, 2, \ldots, k$, also have a further role. They provide the k contributions with $(n_i - 1)$ degrees of freedom, $i = 1, 2, \ldots, k$, respectively, that must be added together to give the pure error sum of squares.

In summary, then, we can still perform the basic regression calculations if only the means and sample variance estimates are provided for each set of repeat runs. We do not need to know the individual observations themselves.

Note. The above comments will help you work Exercise G.

EXERCISES FOR CHAPTER 22

A. Here is a suggestion for a project. Consult, for example, company annual reports, or look up the same information in publications available at many libraries. Obtain the annual earnings per share \$$Y_i$ for several consecutive years for one or more companies, your choice. For each company, fit a polynomial model $Y_i = f(\text{Year}) + \epsilon$ of suitable order that will explain the data satisfactorily, using the method of orthogonal polynomials.

Investors like to see the companies whose shares they own increase earnings per share at a satisfactory *rate* measured in terms of percentage (or proportional) increase per year *over the previous year's results*. For this reason, it makes sense, provided earnings are positive, to examine not only earnings per share, but also the logarithms of these earnings. If earnings increased at a constant rate, for example, log(earnings) would exhibit a straight line trend. Take logarithms to the base e, or to the base 10 (the results differ only by a constant factor), of the data you selected and use the method of orthogonal polynomials to fit a polynomial of suitable order to the results. Compare, for each individual company you examine, the results of the fits to the unlogged and logged data. Check the residuals to see if the error assumptions appear to be violated. Comment on your results.

The following example data are genuine, but are given anonymously:

Company A: 1.93, 0.20, 2.55, 4.31, 8.51, 9.20, 5.10, 3.46, 0.99, 2.33, 3.37.

Company B: 2.17, −0.43, −0.72, −1.81, 5.48, 4.46, 4.37, −6.09, −9.29, 2.44, 2.20.

B. Fit a cubic equation using orthogonal polynomials to the Y-values 13, 4, 3, 4, 10, 22, which are equally spaced in the respective X-values given by $2X = -5, -3, -1, 1, 3, 5$, that is, $X = -2.5, -1.5, \ldots$. Is the cubic term needed? If not, what is the best quadratic, both in orthogonal polynomial form and in terms of X?

If the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$ had been fitted directly, how would the extra sums of squares $SS(b_1|b_0) = 58.51$, $SS(b_2|b_0, b_1) = 210.58$, $SS(b_3|b_0, b_1, b_2) = 0.006$ relate to the sums of squares for the first-, second-, and third-order orthogonal polynomials?

C. A newly born baby was weighed weekly, the figure adopted in each case being the average of the weights on three successive days. Twenty such weights are shown below, recorded in ounces. Fit to the data, using orthogonal polynomials, a polynomial model of a degree justified by the accuracy of the figures; that is, test as you go along for the significance of the linear, quadratic, and so forth, terms.

Number of week	1	2	3	4	5	6	7	8	9	10
Weight	141	144	148	150	158	161	166	170	175	181
Number of week	11	12	13	14	15	16	17	18	19	20
Weight	189	194	196	206	218	229	234	242	247	257

Source: Cambridge Diploma, 1950. Published with permission of Cambridge University Press.

D. A finished product is known to lose weight after it is produced. The following data demonstrate this drop in weight.

Time After Production, t	Weight Difference (in $\frac{1}{16}$ oz), Y
0	0.21
0.5	−1.46
1.0	−3.04
1.5	−3.21
2.0	−5.04
2.5	−5.37
3.0	−6.03
3.5	−7.21
4.0	−7.46
4.5	−7.96

Requirements

1. Using orthogonal polynomials, develop a second-order fitted equation that represents the loss in weight as a function of time after production.
2. Analyze the residuals from this model and draw conclusions about its adequacy.

E. Nine equally spaced levels of a dyestuff were applied to apparently identical pieces of cloth. The color ratings awarded, in order of increasing dyestuff levels, were

$$Y = 11,\quad 12,\quad 10,\quad 12,\quad 11,\quad 14,\quad 16,\quad 22,\quad 28.$$

Find a suitable polynomial relationship between Y and the level of dyestuff using orthogonal polynomials.

F. A mother keeps records of the monthly number of ice-cream portions served to her family over a 12-month period, with the results below. She wonders if the pattern can be explained by a polynomial regression model, and asks you to carry out such an analysis using the orthogonal polynomials of sixth and smaller orders. Do this, provide all the standard analyses, as indicated below, and set out the conclusions to which your solution leads.

Month	Portions Served
January	2
February	72
March	106
April	116
May	103
June	82
July	54
August	30
September	14
October	23
November	64
December	129

Details Required

Model fitting

ANOVA table

Decision on model order to use and telescoping of ANOVA, and so on

R^2, F-tests, and so on.

Estimated variance–covariance matrix

Evaluation of residuals *from chosen model*

Estimated variance–covariance matrix of residuals, diagonal elements only, and plots of residuals

Tests of residuals, including Durbin–Watson, and time sequence runs

Figures, as appropriate

Conclusions

G. The data below are selected from "'Health' and 'Ailments' of M.S.U. Alumni: A Study in Retrospect Over a Five Year Period (1961–62 to 1965–66)" by Mrs. J. V. Bhanot and Shri R. C. Patel, published by the Department of Statistics at the Maharaja Sayajirao University of Baroda, in October 1968. The data show average weights in pounds and standard deviations of various categories of students for 1965–1966. Only a portion of the data is reproduced here, and it is not necessarily fully representative of the whole. Because individual data points are not available, treat each set of repeat runs as the same number of all equal runs for regression fitting purposes, but use the corresponding s.d. (standard deviation) figure, squared and multiplied by the corresponding degrees of freedom (namely, number − 1) to obtain a suitable contribution to the pure error sum of squares. Be careful, because you will have to make suitable adjustments to recover proper sums of squares entries elsewhere.

Fit a model of the form $Y = \beta_0 + \beta_1(\text{age}) + \epsilon$ to the data, where Y = weight of student, adding suitable dummy variables to distinguish the different categories of persons. Provide a detailed analysis. What are your conclusions?

If you feel the model fitted is unsuitable, what model would you suggest instead, and why? Fit any alternative model you have suggested, perform the appropriate analysis, and state your conclusions.

Age in Years	Number of Students	Average Weight, Pounds	Est. s.d.	Category[a]
16	19	103.82	12.70	*A*
17	19	105.39	12.50	*A*
18	16	107.50	15.30	*A*
19	8	103.12	7.68	*A*
20	6	105.83	16.75	*A*
21	6	107.50	5.77	*A*
22	1	102.50	—	*A*
23	1	117.50	—	*A*
16	12	99.17	9.20	*B*
17	29	107.67	19.55	*B*
18	28	103.57	13.05	*B*
19	32	112.19	17.60	*B*
20	22	110.00	12.04	*B*
21	8	113.13	14.87	*B*
22	4	112.00	11.18	*B*
23	7	113.21	7.28	*B*
16	18	100.83	16.75	*C*
17	33	99.32	14.50	*C*

Age in Years	Number of Students	Average Weight, Pounds	Est. s.d.	Category[a]
18	24	100.83	19.67	*C*
19	18	96.39	13.29	*C*
20	6	101.67	18.12	*C*
21	4	100.00	15.21	*C*
22	0	—	—	*C*
23	1	112.50	—	*C*
Total	322			

[a] *A* = Hindu eggitarian male students; *B* = non-Hindu mixed (diet) male students; *C* = Hindu mixed (diet) female students.