10.2: Quantitative Predictors - Orthogonal Polynomials

Polynomial trends in the response with respect to a quantitative predictor can be evaluated by using orthogonal polynomial contrasts, a special set of linear contrasts. This is an alternative to the Regression analysis illustrated in the previous section, which may be affected by multicollinearity. Note that centering to remedy multicollinearity is effective only for quadratic polynomials. Therefore, this simple technique of trend analysis performed via orthogonal polynomial coding will prove to be beneficial for higher-order polynomials. Orthogonal polynomials have the property that the cross-products defined by the numerical coefficients of their terms add to zero.

The orthogonal polynomial coding can be applied only when the levels of quantitative predictor are equally spaced. The method is to partition the quantitative factor in the ANOVA table into independent single degrees of freedom comparisons. The comparisons are called orthogonal polynomial contrasts or comparisons.

Orthogonal polynomials are equations such that each is associated with a power of the independent variable (e.g. x, linear; x^2, quadratic; x^3, cubic, etc.). In other words, orthogonal polynomials are coded forms of simple polynomials. The number of possible comparisons is equal to $k-1$, where k is the number of quantitative factor levels. For example, if $k=3$, only two comparisons are possible allowing for testing of linear and quadratic effects.

Using orthogonal polynomials to fit the desired model to the data would allow us to eliminate collinearity and to seek the same information as simply polynomials.

A typical polynomial model of order k would be:

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_k x^k+\epsilon \tag{10.2.1}$$

The simple polynomials used are $x, x^2, \ldots, x^k$. We can obtain orthogonal polynomials as linear combinations of these simple polynomials. If the levels of the predictor variable, x, are equally spaced, then one can easily use coefficient tables to determine the orthogonal polynomial coefficients that can be used to set up an orthogonal polynomial model.

If we are to fit the k^{th} order polynomial to using orthogonal contrasts coefficients, the general equation can be written as

$$y_{ij}=\alpha_0+\alpha_1 g_{1i}(x)+\alpha_2 g_{2i}(x)+\cdots+\alpha_k g_{ki}(x)+\epsilon_{ij} \tag{10.2.2}$$

where $g_{pi}(x)$ is a polynomial in x of degree $p, (p=1,2,\ldots,k)$ for the i^{th} level treatment factor and the parameter α_p depends on the coefficients β_p. Using the properties of the function $g_{pi}(x)$, one can show that the first five orthogonal polynomial are of the following form:

$$\text{Mean:} \quad g_0(x)=1 \tag{10.2.3}$$

$$\text{Linear:} \quad g_1(x)=\lambda_1\left(\frac{x-\bar{x}}{d}\right) \tag{10.2.4}$$

$$\text{Quadratic:} \quad g_2(x)=\lambda_2\left(\left(\frac{x-\bar{x}}{d}\right)^2-\left(\frac{t^2-1}{12}\right)\right) \tag{10.2.5}$$

$$\text{Cubic:} \quad g_3(x)=\lambda_3\left(\left(\frac{x-\bar{x}}{d}\right)^3-\left(\frac{x-\bar{x}}{d}\right)\left(\frac{3t^2-7}{20}\right)\right) \tag{10.2.6}$$

$$\text{Quartic:} \quad g_4(x)=\lambda_4\left(\left(\frac{x-\bar{x}}{d}\right)^4-\left(\frac{x-\bar{x}}{d}\right)^2\left(\frac{3t^2-13}{14}\right)+\frac{3\left(t^2-1\right)\left(t^2-9\right)}{560}\right) \tag{10.2.7}$$

where t = number of levels of the factor, x = value of the factor level, $\bar{x}$ = mean of the factor levels, and d = distance between factor levels.

In the next section, we will illustrate how the orthogonal polynomial contrast coefficients are generated, and the Factor SS is partitioned. This method will be required to fit polynomial regression models with terms greater than the quadratic, because even after centering there will still be multicollinearity between x and x^3 as well as between x^2 and x^4.

The following example is taken from *Design of Experiments: Statistical Principles of Research Design and Analysis* by Robert Kuehl.

Example 10.2.1: Grain Yield

The treatment design consisted of five plant densities (10, 20, 30, 40, and 50). Each of the five treatments was assigned randomly to three field plots in a completely randomized experimental design. The resulting grain yields are shown in the table below (Grain Data):

	Plant Density (x)				
	10	20	30	40	50
	12.2	16.0	18.6	17.6	18.0
	11.4	15.5	20.2	19.3	16.4
	12.4	16.5	18.2	17.1	16.6
Means $(\bar{y}_i)$	12.0	16.0	19.0	18.0	17.0

Solution

We can see that the factor levels of plant density are equally spaced. Therefore, we can use the orthogonal contrast coefficients to fit a polynomial to the response, grain yields. With $k = 5$, we can only fit up to a quartic term. The orthogonal polynomial contrast coefficients for the example are shown in Table 10.1.

Table 10.1 - Computations for orthogonal polynomial contrasts and sums of squares

Density (x)	$\bar{y}_i$	Orthogonal Polynomial Coefficients (g_{pi})				
		Mean	Linear	Quadratic	Cubic	Quartic
10	12	1	-2	2	-1	1
20	16	1	-1	-1	2	-4
30	19	1	0	-2	0	6
40	18	1	1	-1	-2	-4
50	17	1	2	2	1	1
λ_p		-	1	1	5/6	35/12
Sum = $\sum g_{pi}\bar{y}_i$		82	12	-14	1	7
Divisor = $\sum g_{pi}^2$		5	10	14	10	70
$SSP_p = r(\sum g_{pi}\bar{y}_i)^2 / \sum g_{pi}^2$		-	43.2	42.0	0.3	2.1
$\hat{\alpha}_p = \sum g_{pi}\bar{y}_i / \sum g_{pi}^2$		16.4	1.2	-1.0	0.1	0.1

As mentioned before, one can easily find the orthogonal polynomial coefficients for a different order of polynomials using pre-documented tables for equally spaced intervals. However, let us try to understand how the coefficients are obtained.

First note that the five values of x are $10, 20, 30, 40, 50$ Therefore, $\bar{x} = 30$ and the spacing $d = 10$. This means that the five values of $\frac{x-\bar{x}}{d}$ are $-2, -1, 0, 1,$ and 2.

Linear coefficients: The polynomial g_1 for linear coefficients turn out to be:

Linear Coefficient Polynomials g_1					
x	10	20	30	40	50
$(x-30)$	-20	-10	0	10	20
$\frac{(x-30)}{10}$	-2	-1	0	1	2
Linear orthogonal polynomial	$(-2)\lambda_1$	$(-1)\lambda_1$	$(0)\lambda_1$	$(1)\lambda_1$	$(2)\lambda_1$

To obtain the final set of coefficients we choose λ_1 so that the coefficients are integers. Therefore, we set $\lambda_1 = 1$ and obtain the coefficient values in Table 10.1.

Quadratic coefficients: The polynomial g_2 for linear coefficients:

Linear Coefficient Polynomials g_2					
Linear orthogonal polynomial	$\left((-2)^2 - \left(\frac{5^2-1}{12}\right)\right)\lambda_2$	$\left((-1)^2 - \left(\frac{5^2-1}{12}\right)\right)\lambda_2$	$\left((0)^2 - \left(\frac{5^2-1}{12}\right)\right)\lambda_2$	$\left((1)^2 - \left(\frac{5^2-1}{12}\right)\right)\lambda_2$	$\left((2)^2 - \left(\frac{5^2-1}{12}\right)\right)\lambda_2$
Simplified form	$(2)\lambda_2$	$(-1)\lambda_2$	$(-2)\lambda_2$	$(-1)\lambda_2$	$(2)\lambda_2$

To obtain the final set of coefficients we choose λ_2 so that the coefficients are integers. Therefore, we set $\lambda_2 = 1$ and obtain the coefficient values in Table 10.1.

Cubic coefficients: The polynomial g_3 for linear coefficients:

Linear Coefficient Polynomials g_3					
Linear orthogonal polynomial	$\left((-2)^3 - (-2)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3$	$\left((-1)^3 - (-1)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3$	$\left((0)^3 - (0)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3$	$\left((1)^3 - (1)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3$	$\left((2)^3 - (2)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3$
Simplified form	$\left(-\frac{6}{5}\right)\lambda_3$	$\left(\frac{12}{5}\right)\lambda_3$	$(0)\lambda_3$	$\left(-\frac{12}{5}\right)\lambda_3$	$\left(\frac{6}{5}\right)\lambda_3$

Quartic coefficients: The polynomial g4 can be used to obtain the quartic coefficients in the same way as above.

Notice that each set of coefficients for contrast among the treatments since the sum of coefficients is equal to zero. For example, the quartic coefficients $(1, -4, 6, -4, 1)$ sums to zero. Using orthogonal polynomial contrasts, we can partition the treatment sums of squares into a set of additive sums of squares corresponding to orthogonal polynomial contrasts. Computations are similar to what we learned in lesson 2.5. We can use those partitions to test sequentially the significance of linear, quadratic, cubic, and quartic terms in the model to find the polynomial order appropriate for the data.

Table 10.1 shows how to obtain the sums of squares for each component and how to compute the estimates of the α_p coefficients for the orthogonal polynomial equation. Using the results in table 10.1, we have estimated orthogonal polynomial equation as:

$$\hat{y}_i = 16.4 + 1.2g_{1i} - 1.0g_{2i} + 0.1g_{3i} + 0.1g_{4i}$$

Table 10.2 summarizes how the treatment sums of squares are partitioned and their test results.

Table 10.2 - Analysis of variance for the orthogonal polynomial model relationship between plant density and grain yield.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	Pr > F
Density	4	87.60	21.90	29.28	F">.000
Error	10	7.48	0.75		F">

Contrast	DF	Contrast SS	Mean Square	F	Pr > F
Linear	1	43.20	43.20	57.75	F">.000
Quadratic	1	42.00	42.00	56.15	F">.000
Cubic	1	.30	.30	.40	F">.541
Quartic	1	2.10	2.10	2.81	F">.125

To test whether any of the polynomials are significant (i.e. $H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$), we can use the global F-test where the test statistic is equal to 29.28. We see that the p-value is almost zero and therefore we can conclude that at the 5% level at least one of the polynomials is significant. Using the orthogonal polynomial contrasts we can determine which of the polynomials are useful. From table 3.5, we see that for this example only the linear and quadratic terms are useful. Therefore we can write the estimated orthogonal polynomial equation as:

$$16.4 + 1.2g_{1i} - 1.0g_{2i}$$

The polynomial relationship expressed as a function of y and x in actual units of the observed variables is more informative than when expressed in units of the orthogonal polynomial.

We can obtain the polynomial relationship using the actual units of observed variables by back-transforming using the relationships presented earlier. The necessary quantities to back-transform are $\lambda_1 = 1$, $d = 10$, $\bar{x} = 30$, and $t = 5$. Substituting these values, we obtain

$$\begin{aligned}\hat{y} &= 16.4 + 1.2g_{1i} - 1.0g_{2i} \\ &= 16.4 + 1.2(1)\left(\frac{x-30}{10}\right) - 1.0(1)\left(\left(\frac{x-30}{10}\right)^2 - \frac{5^2-1}{12}\right)\end{aligned}$$

which simplifies to

$$\hat{y} = 5.8 + 0.72x - 0.01x^2$$

Generating Orthogonal Polynomials

? Using SAS

Steps in SAS

Below is the code for generating polynomials from the IML procedure in SAS:

```
/* read the grain data set */
/* Generating Ortho_Polynomials from IML */
proc iml;
x={10 20 30 40 50};
xpoly=orpol(x,4); /* the '4' is the df for the quantitative factor */
density=x`; new=density || xpoly;
create out1 from new[colname={"density" "xp0" "xp1" "xp2" "xp3" "xp4"}];
append from new; close out1;
quit;
proc print data=out1;
run;
/* Here data is sorted and then merged with the original dataset */
proc sort data=grain;
by density;
run;
```

```
data ortho_poly; merge out1 grain;
by density;
run;
proc print data=ortho_poly;
run;
/* The following code will then generate the results shown in the
Online Lesson Notes for the Kuehl example data */
proc mixed data=ortho_poly method=type3;
class;
model yield=xp1 xp2 xp3 xp4;
title 'Using Orthog polynomials from IML';
run;
/* We can use proc glm to obtain the same results without using
IML codings, to directly obtained the same results.
Proc glm will use the orthogonal contrast coefficients directly */
proc glm data=grain;
class density;
model yield = density;
contrast 'linear' density -2 -1 0 1 2;
contrast 'quadratic' density 2 -1 -2 -1 2;
contrast 'cubic' density -1 2 0 -2 1;
contrast 'quartic' density 1 -4 6 -4 1;
run;
```

The output is:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
xp1	1	43.200000	43.200000	Var(Residual) + Q(xp1)	MS(Residual)	10	57.75	F"> < .0001
xp2	1	42.000000	42.000000	Var(Residual) + Q(xp2)	MS(Residual)	10	56.15	F"> < .0001
xp3	1	0.300000	0.300000	Var(Residual) + Q(xp2)	MS(Residual)	10	0.40	F"> 0.5407
xp4	1	2.100000	2.100000	Var(Residual) + Q(xp4)	MS(Residual)	10	2.81	F"> 0.1248
Residual	10	7.480000	7.480000	Var(Residual)				F" class=" ">

Vari…
Vari…
Vari…
Vari…

Fitting a Quadratic Model with Proc Mixed

Often we can see that only a quadratic curvature is of interest in a set of data. In this case, we can plan to simply run an order 2 (quadratic) polynomial and can easily use proc mixed (the general linear model). This method just requires centering the quantitative variable levels by subtracting the mean of the levels (30) and then creating the quadratic polynomial terms.

```
data grain;
set grain;
x=density-30;
x2=x**2;
run;
proc mixed data=grain method=type3;
class;
model yield = x x2;
run;
```

The output is:

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
x	1	43.200000	43.200000	Var(Residual) + Q(x)	MS(residual)	12	52.47	F"> <.0001
x2	1	42.000000	42.000000	Var(Residual) + Q(x2)	MS(residual)	12	51.01	F"> <.0001
Residual	12	9.880000	0.823333	Var(Residual)				F" class=" ">

We can also generate the solutions (coefficients) for the model with:

```
proc mixed data=grain method=type3;
class;
model yield = x x2 / solution;
run;
```

which gives the following output for the regression coefficients:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	18.4000	0.3651	12	50.40	|t"> <.0001
x	0.1200	0.01657	12	7.24	|t"> <.0001
x2	-0.01000	0.001400	12	-7.14	|t"> <.0001

Here we need to keep in mind that the regression was based on centered values for the predictor, so we have to back-transform to get the coefficients in terms of the original variables. This back-transform process (from Kutner et.al) is:

Regression Function in Terms of X

After a polynomial regression model has been developed, we often wish to express the final model in terms of the original variables rather than keeping it in terms of the centered variables. This can be done readily. For example, the fitted second-order model for one predictor variable that is is expressed in terms of centered values $x = X - \bar{X}$:

$$\hat{Y} = b_0 + b_1(x) + b_{11}x^2$$

because in terms of the original X variable:

$$\hat{Y} = b'_0 + b'_1 X + b'_{11} X^2$$

where:

$$\begin{aligned} b'_0 &= b_0 - b_1 \bar{X} + b_{11} \bar{X}^2 \\ b'_1 &= b_1 - 2 b_{11} \bar{X} \\ b'_{11} &= b_{11} \end{aligned}$$

In the example above, this back-transformation uses the estimates from the Solutions for Fixed Effects table above.

```
data backtransform;
bprime0=18.4-(.12*30)+(-.01*(30**2));
bprime1=.12-(2*-.01*30);
bprime2=-.01;
title 'bprime0=b0-(b1*meanX)+(b2*(meanX)2)';
title2 'bprime1=b1=2*b2*meanX';
title3 'bprime2=b2';
run;
proc print data=backtransform;
var bprime0 bprime1 bprime2;
run;
```

The output is then:

Obs	bprime0	bprime1	bprime2
1	5.8	0.72	-0.01

Note

The ANOVA results and the final quadratic regression equation here are identical to the results from the orthogonal polynomial coding approach.

Using R

- Load the Grain Data.
- Obtain the ANOVA table.
- Fit a quadratic model after centering the covariate and creating x^2. Transform back to the original variables.

Steps in R

1. Load the Grain data and obtain the ANOVA table by using the following commands:

```
setwd("~/path-to-folder/")
grain_data <- read.table("grain_data.txt",header=T)
attach(grain_data)
poly_model<-lm(yield ~ poly(density,4),data=grain_data)
summary(poly_model)
#Coefficients:
#                  Estimate  Std. Error  t value  Pr(>|t|)
#(Intercept)         16.4000      0.2233   73.441  5.35e-15 ***
```

```
#poly(density, 4)1    6.5727      0.8649    7.600  1.84e-05 ***
#poly(density, 4)2   -6.4807      0.8649   -7.493  2.08e-05 ***
#poly(density, 4)3    0.5477      0.8649    0.633     0.541
#poly(density, 4)4    1.4491      0.8649    1.676     0.125
anova(poly_model)
#Analysis of Variance Table
#Response: yield
#                  Df  Sum Sq  Mean Sq  F value    Pr(>F)
#poly(density, 4)   4   87.60   21.900   29.278  1.69e-05 ***
#Residuals         10    7.48    0.748
#---
#Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

By using the command `anova()` we can test whether any of the polynomials are significant (i.e. $H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$. We can use the global F-test where the test statistic is equal to 29.28. We see that the p-value is almost zero, and therefore we can conclude that at the 5% level at least one of the polynomials is significant.

By using the command `summary()` we can test which contrasts are significant. For this example only the linear and quadratic terms are significant since there p-values are almost zero.

2. Fit a quadratic model after centering the covariate and creating x^2 by using the following commands:

Transform back to the original variables

```
density_center<-density-30
density_square_center<-density_center^2
new_data<-cbind(grain_data,density_center,density_square_center)
ancova_model<-lm(yield ~ density_center + density_square_center,new_data)
summary(ancova_model)
#Coefficients:
#                      Estimate  Std. Error  t value  Pr(>|t|)
#(Intercept)           18.40000     0.36511   50.396  2.44e-15 ***
#density_center         0.12000     0.01657    7.244  1.02e-05 ***
#density_square_center -0.01000     0.00140   -7.142  1.18e-05 ***
#---
#Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
anova(ancova_model)
#Analysis of Variance Table
#Response: yield
#                      Df  Sum Sq  Mean Sq  F value     Pr(>F)
#density_center         1   43.20   43.200   52.470  1.024e-05 ***
#density_square_center  1   42.00   42.000   51.012  1.177e-05 ***
#Residuals             12    9.88    0.823
#---
#Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. Transform back to the original variables

The estimated coefficients for the polynomial model are 18.4, 0.12 and -0.01. Here we need to keep in mind that the regression was based on centered values for the predictor, so we have to back-transform to get the coefficients in terms of the original variables. We can do that by using the following commands:

```
b_0_prime<-18.4-0.12*30-0.01*30^2 #5.8
b_1_prime<-0.12-0.01*(-2*30) # 0.72
b_2_prime<--0.01 # -0.01
detach(grain_data)
```

For the original variables the estimated coefficients are 5.8, 0.72 and -0.01.

This page titled 10.2: Quantitative Predictors - Orthogonal Polynomials is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Penn State's Department of Statistics via source content that was edited to the style and standards of the LibreTexts platform.