10.2 - Prédicteurs quantitatifs : polynômes orthogonaux

10.2 - Prédicteurs quantitatifs : polynômes orthogonaux

Les tendances polynomiales de la réponse par rapport à un prédicteur quantitatif peuvent être évaluées en utilisant des contrastes polynomiaux orthogonaux, un ensemble spécial de contrastes linéaires. Il s'agit d'une alternative à l'analyse de régression illustrée dans la section précédente, qui peut être affectée par la multicolinéarité. Notez que le centrage pour remédier à la multicolinéarité n'est efficace que pour les polynômes quadratiques. Par conséquent, cette technique simple d'analyse de tendance effectuée via un codage polynomial orthogonal s'avérera bénéfique pour les polynômes d'ordre supérieur. Les polynômes orthogonaux ont la propriété que les produits vectoriels définis par les coefficients numériques de leurs termes s'additionnent à zéro.

Le codage polynomial orthogonal ne peut être appliqué que lorsque les niveaux de prédicteur quantitatif sont équidistants. La méthode consiste à partitionner le facteur quantitatif dans le tableau ANOVA en comparaisons indépendantes à un seul degré de liberté. Les comparaisons sont appelées contrastes polynomiaux orthogonaux ou comparaisons.

Les polynômes orthogonaux sont des équations telles que chacune est associée à une puissance de la variable indépendante (par exemple x, linéaire ;x^2, quadratique ;x^3, cubique, etc.). En d'autres termes, les polynômes orthogonaux sont des formes codées de polynômes simples. Le nombre de comparaisons possibles est égal à$k-1$, où$k$est le nombre de niveaux de facteurs quantitatifs. Par exemple, si$k=3$, seules deux comparaisons sont possibles permettant de tester les effets linéaires et quadratiques.

L'utilisation de polynômes orthogonaux pour ajuster le modèle souhaité aux données nous permettrait d'éliminer la colinéarité et de rechercher les mêmes informations que de simples polynômes.

Un modèle polynomial typique d'ordre$k$serait:

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_k x^k+\epsilon$$

Les polynômes simples utilisés sont$x, x^2, \ldots, x^k$. Nous pouvons obtenir des polynômes orthogonaux comme combinaisons linéaires de ces polynômes simples. Si les niveaux de la variable prédictive,$x$sont équidistants, on peut facilement utiliser des tables de coefficients pour déterminer les coefficients polynomiaux orthogonaux qui peuvent être utilisés pour établir un modèle polynomial orthogonal.

Si nous devons adapter lek^{th}polynôme d'ordre à l'aide de coefficients de contrastes orthogonaux, l'équation générale peut être écrite comme

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

où$g_{pi}(x)$est un polynôme en$x$de degré$p, (p=1,2,\ldots,k)$pour lei^{th}facteur de traitement de niveau et le paramètre$\alpha_p$dépend des coefficientsβ_p. Utilisation des propriétés de la fonction $g_{pi}(x)$, on peut montrer que les cinq premiers polynômes orthogonaux sont de la forme suivante :

Moyenne:$g_0(x)=1$

Linéaire:$g_1(x)=\lambda_1\left(\frac{x-\bar{x}}{d}\right)$

Quadratique:$g_2(x)=\lambda_2\left(\left(\frac{x-\bar{x}}{d}\right)^2-\left(\frac{t^2-1}{12}\right)\right)$

Cubique:$g_3(x)=\lambda_3\left(\left(\frac{x-\bar{x}}{d}\right)^3-\left(\frac{x-\bar{x}}{d}\right)\left(\frac{3t^2-7}{20}\right)\right)$

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

oùt= nombre de niveaux du facteur,x= valeur du niveau du facteur,$\bar{x}$= moyenne des niveaux de facteurs, et$d$= distance entre les niveaux de facteurs.

Dans la section suivante, nous illustrerons comment les coefficients de contraste polynomiaux orthogonaux sont générés et comment le facteur SS est partitionné. Cette méthode sera nécessaire pour ajuster les modèles de régression polynomiale avec des termes supérieurs au quadratique, car même après le centrage, il y aura toujours une multicolinéarité entre$x$etx^3ainsi qu'entrex^2etx^4.

L'exemple suivant est tiré de *Design of Experiments: Statistical Principles of Research Design and Analysis* de Robert Kuehl.

Exemple 10-1 : Rendement en grain

La conception du traitement consistait en cinq densités de plantes (10, 20, 30, 40 et 50). Chacun des cinq traitements a été assigné au hasard à trois parcelles de terrain dans une conception expérimentale complètement aléatoire. Les rendements en grains qui en résultent sont indiqués dans le tableau ci-dessous (Grain Data) : [1]

	Densité de plantation (x)				
	dix	**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
Signifie ($\bar{y}_i$)	12.0	16.0	19.0	18.0	17.0

Nous pouvons voir que les niveaux de facteurs de la densité des plantes sont également espacés. Par conséquent, nous pouvons utiliser les coefficients de contraste orthogonaux pour ajuster un polynôme à la réponse, les rendements en grain. Avec$k = 5$, nous ne pouvons nous adapter qu'à un terme quartique. Les coefficients de contraste polynomial orthogonal pour l'exemple sont présentés dans le tableau 10.1.

Tableau 10.1 - Calculs pour les contrastes polynomiaux orthogonaux et les sommes des carrés

Densité(x)	$\bar{y}_i$	Coefficients polynomiaux orthogonaux(g_{pi})				
		Moyenne	**Linéaire**	**Quadratique**	**Cubique**	**quartique**
dix	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
Somme =$\sum g_{pi}\bar{y}_{i.}$		82	12	-14	1	sept
Diviseur =$\sum g_{pi}^2$		5	dix	14	dix	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

Comme mentionné précédemment, on peut facilement trouver les coefficients polynomiaux orthogonaux pour un ordre différent de polynômes en utilisant des tables pré-documentées pour des intervalles équidistants. Cependant, essayons de comprendre comment les coefficients sont obtenus.

Notons tout d'abord que les cinq valeurs de$x$sont 10, 20, 30, 40, 50. Par conséquent,$\bar{x} = 30$et l'espacement$d = 10$. Cela signifie que les cinq valeurs de$\frac{x - \bar{x}}{d}$sont -2, -1, 0, 1 et 2.

Coefficients linéaires : le polynôme$g_1$pour les coefficients linéaires s'avèrent être :

Polynômes de coefficients linéaires					
x	dix	20	30	40	50
$(x - 30)$	-20	-dix	0	dix	20
$\frac{(x-30)}{10}$	-2	-1	0	1	2
Polynôme orthogonal linéaire	$(-2)\lambda_1$	$(-1)\lambda_1$	$(0)\lambda_1$	$(1)\lambda_1$	$(2)\lambda_1$

Pour obtenir l'ensemble final de coefficients, nous choisissons$\lambda_1$de sorte que les coefficients sont des nombres entiers. Par conséquent, nous fixons$\lambda_1 = 1$et obtenir les valeurs des coefficients dans le tableau 10.1.

Coefficients quadratiques : Le polynôme$g_2$pour les coefficients linéaires s'avèrent être :

$$\left((-2)^2-\left(\frac{5^2-1}{12}\right)\right)\lambda_2, \left((-1)^2-\left(\frac{5^2-1}{12}\right)\right)\lambda_2,$$
$$\left((0)^2-\left(\frac{5^2-1}{12}\right)\right)\lambda_2, \left((1)^2-\left(\frac{5^2-1}{12}\right)\right)\lambda_2,$$
$$\left((2)^2-\left(\frac{5^2-1}{12}\right)\right)\lambda_2$$

qui simplifié donne :

$$(2)\lambda_2, (-1)\lambda_2, (-2)\lambda_2, (-1)\lambda_2, (2)\lambda_2$$

Pour obtenir l'ensemble final de coefficients, nous choisissons$\lambda_2$de sorte que les coefficients sont des nombres entiers. Par conséquent, nous fixons$\lambda_2 = 1$et obtenir les valeurs des coefficients dans le tableau 10.1.

Coefficients cubiques : Le polynôme$g_3$pour les coefficients cubiques s'avèrent être:

$$\left((-2)^3-(-2)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3, \left((-1)^3-(-1)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3,$$
$$\left((0)^3-(0)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3, \left((1)^3-(1)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3,$$
$$\left((2)^3-(2)\left(\frac{3(5^2)-7}{20}\right)\right)\lambda_3$$

qui simplifié donne :

$$(-\frac{6}{5})\lambda_3, (\frac{12}{5})\lambda_3, (0)\lambda_3, (\frac{-12}{5})\lambda_3, (\frac{6}{5})\lambda_3$$

Pour obtenir l'ensemble final de coefficients, nous choisissonsλ_3 de sorte que les coefficients sont des nombres entiers. Par conséquent, nous fixons$\lambda_3 = 5/6$et obtenir les valeurs des coefficients dans le tableau 10.1.

Coefficients quartiques : le polynôme$g_4$peut être utilisé pour obtenir les coefficients quartiques de la même manière que ci-dessus.

Notez que chaque ensemble de coefficients de contraste entre les traitements puisque la somme des coefficients est égale à zéro. Par exemple, les coefficients quartiques$(1, -4, 6, -4, 1)$sommes à zéro. En utilisant des contrastes de polynômes orthogonaux, nous pouvons partitionner les sommes de carrés de traitement en un ensemble de sommes de carrés additives correspondant à des contrastes de polynômes orthogonaux. Les calculs sont similaires à ce que nous avons appris dans la leçon 2.5 . Nous pouvons utiliser ces partitions pour tester séquentiellement la signification des termes linéaires, quadratiques, cubiques et quartiques dans le modèle afin de trouver l'ordre polynomial approprié pour les données. [2]

Le tableau 10.1 montre comment obtenir les sommes des carrés pour chaque composante et comment calculer les estimations desα_p coefficients pour l'équation polynomiale orthogonale. En utilisant les résultats du tableau 10.1, nous avons estimé l'équation polynomiale orthogonale comme suit :

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

Le tableau 10.2 résume la répartition des sommes des carrés de traitement et leurs résultats de test.

Tableau 10.2 - Analyse de variance pour la relation du modèle polynomial orthogonal entre la densité de peuplement et le rendement en grains.

Source de variation	Degrés de liberté	Somme des carrés	Carré moyen	F	Pr > F
Densité	4	87,60	21.90	29.28	.000
Erreur	dix	7.48	0,75		

Contraste	DF	Contraste SS	Carré moyen	F	Pr > F
Linéaire	1	43.20	43.20	57,75	.000
Quadratique	1	42.00	42.00	56.15	.000
Cubique	1	.30	.30	.40	.541
quartique	1	2.10	2.10	2,81	.125

Pour tester si l'un des polynômes est significatif (c'est-à-dire$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, nous pouvons utiliser le test F global où la statistique de test est égale à 29,28. Nous voyons que la valeur de p est presque nulle et nous pouvons donc conclure qu'au niveau de 5 % au moins un des polynômes est significatif. En utilisant les contrastes polynomiaux orthogonaux, nous pouvons déterminer lesquels des

polynômes sont utiles. Du tableau 3.5, on voit que pour cet exemple seuls les termes linéaires et quadratiques sont utiles. Par conséquent, nous pouvons écrire l'équation polynomiale orthogonale estimée sous la forme :

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

La relation polynomiale exprimée en fonction de *y* et *x* en unités réelles des variables observées est plus informative que lorsqu'elle est exprimée en unités du polynôme orthogonal.

Nous pouvons obtenir la relation polynomiale en utilisant les unités réelles des variables observées en rétro-transformant en utilisant les relations présentées précédemment. Les quantités nécessaires pour rétro-transformer sont$\lambda_1 = 1, d = 10, \bar{x} = 30$et *t* = 5. En remplaçant ces valeurs, nous obtenons

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

qui se simplifie en

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

SAS® avec SAS

Génération de polynômes orthogonaux dans SAS

Vous trouverez ci-dessous le code permettant de générer des polynômes à partir de la procédure IML dans 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 'quatic' density 1 -4 6 -4 1;
run;
```

Analyse de la variance								
La source	**DF**	**Somme des carrés**	**Carré moyen**	**Carré moyen attendu**	**Terme d'erreur**	**Erreur DF**	**Valeur F**	**Pr > F**
xp1	1	43.200000	43.200000	Var(Résiduel) + Q(xp1)	MS (résiduel)	dix	57,75	< 0,0001
xp2	1	42.000000	42.000000	Var(Résiduel) + Q(xp2)	MS (résiduel)	dix	56.15	< 0,0001
xp3	1	0,300000	0,300000	Var(Résiduel) + Q(xp2)	MS (résiduel)	dix	0,40	0,5407
xp4	1	2.100000	2.100000	Var(Résiduel) + Q(xp4)	MS (résiduel)	dix	2,81	0,1248
Résiduel	dix	7,480000	7,480000	Var(Résiduel)				

Ajustement d'un modèle quadratique avec Proc Mixed

Souvent, nous pouvons voir que seule une courbure quadratique est intéressante dans un ensemble de données. Dans ce cas, nous pouvons envisager d'exécuter simplement un polynôme d'ordre 2 (quadratique) et pouvons facilement utiliser proc mixed (le modèle linéaire général). Cette méthode nécessite simplement de centrer les niveaux des variables quantitatives en soustrayant la moyenne des niveaux (30) puis de créer les termes polynomiaux quadratiques.

```
data grain;
  set grain;
  x=density-30;
  x2=x**2;
run;

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

La sortie est :

Analyse de variance de type 3								
La source	**DF**	**Somme des carrés**	**Carré moyen**	**Carré moyen attendu**	**Terme d'erreur**	**Erreur DF**	**Valeur F**	**Pr > F**
X	1	43.200000	43.200000	Var(Résiduel) + Q(x)	MS (résiduel)	12	52,47	<.0001
x2	1	42.000000	42.000000	Var(Résiduel) + Q(x2)	MS (résiduel)	12	51.01	<.0001
Résiduel	12	9.880000	0,823333	Var(Résiduel)				

On peut aussi générer les solutions (coefficients) du modèle avec :

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

qui donne la sortie suivante pour les coefficients de régression :

Solution pour les effets fixes					
Effet	**Estimation**	**Erreur standard**	**DF**	**t Valeur**	**Pr > |t|**
Intercepter	18.4000	0,3651	12	50,40	<.0001
X	0,1200	0,01657	12	7.24	<.0001
x2	-0.01000	0,001400	12	-7.14	<.0001

Ici, nous devons garder à l'esprit que la régression était basée sur des valeurs centrées pour le prédicteur, nous devons donc effectuer une rétrotransformation pour obtenir les coefficients en fonction des variables d'origine. Ce processus de rétro-transformation (de Kutner et.al) est :

Fonction de régression en termes de *X* . Une fois qu'un modèle de régression polynomial a été développé, nous souhaitons souvent exprimer le modèle final en fonction des variables d'origine plutôt que de le conserver en fonction des variables centrées. Cela peut être fait facilement. Par exemple, le modèle de second ordre ajusté pour une variable prédictive exprimée en termes de valeurs centrées $x = X - \bar{X}$:

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

devient en fonction de la variable *X d'origine :*

$$\hat{Y} = b_0' + b_1' X + b_{11}' X^2$$

où:

$$b_0' = b_0 - b_1 \bar{X} + b_{11} \bar{X}^2$$
$$b_1' = b_1 - 2 b_{11} \bar{X}$$
$$b_{11}' = b_{11}$$

Dans l'exemple ci-dessus, cette rétro-transformation utilise les estimations du tableau Solutions pour les effets fixes ci-dessus.

```
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;
```

La sortie est alors :

Obs	bprime0	bprime1	bprime2
1	5.8	0,72	-0,01

Noter! Les résultats de l'ANOVA et l'équation de régression quadratique finale ici sont identiques aux résultats de l'approche de codage polynomial orthogonal.

Légende

[1]	Lien
↥	A une info-bulle/un popover
⬚	Visibilité commutable

Source : https://online.stat.psu.edu/stat502_fa21/lesson/10/10.2

Liens:

1. https://online.stat.psu.edu/onlinecourses/sites/stat502/files/lesson11/grain_data.txt
2. https://online.stat.psu.edu/stat502_fa21/lesson/2/2.5