

10

Model selection I: principles of model choice and designed experiments

Section 10.1 and the early part of section 10.2 use commands introduced in chapter nine. We therefore move on to polynomials.

10.2 Three principles of model choice

Model choice in the polynomial problem

In the early sections of this chapter, a polynomial model was built up by fitting the three models $Y = X$, $Y = X + X^2$, (or in SAS shorthand $Y = X | X$) and $Y = X + X^2 + X^3$ (or $Y = X | X | X$). Choices can be made between these models on the basis of economy of variables, multiplicity of p-values and considerations of marginality. When considering adjusted sums of squares, only the final p-value in each model is valid. When fitting models for which such considerations of marginality come into play, it is more efficient and informative to employ the tests of the sequential sums of squares, as is shown below for the *simple polynomial* dataset. These are provided automatically in the command line route, but need to be asked for in the menu route.

SAS COMMANDS FOR BOX 10.5 ANOVA tables using sequential rather than adjusted sums of squares

Commands `proc glm data=gandh.Chapter10;`
`model Y1 = X1 | X1 | X1;`
`run;`

Menu route Statistics > Anova > Linear Models...

Y1 → Dependent

X1 → Quantitative

Select 3 as polynomial order by clicking on ▲

Independent → Effects in model

X1 → Polynomial

Type 1

SAS OUTPUT FOR BOX 10.5 ANOVA tables using sequential rather than adjusted sums of squares

The GLM Procedure

Dependent Variable: Y1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6919889.186	2306629.729	379.38	<.0001
Error	76	462077.908	6079.972		
Corrected Total	79	7381967.094			

R-Square	Coeff Var	Root MSE	Y1 Mean
0.937405	25.44390	77.97418	306.4553

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	6663020.646	6663020.646	1095.90	<.0001
X1*X1	1	256148.401	256148.401	42.13	<.0001
X1*X1*X1	1	720.139	720.139	0.12	0.7317

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	2505.485505	2505.485505	0.41	0.5228
X1*X1	1	4762.722850	4762.722850	0.78	0.3789
X1*X1*X1	1	720.138776	720.138776	0.12	0.7317

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-15.74832379	33.92445347	-0.46	0.6438
X1	6.17873499	9.62508176	0.64	0.5228
X1*X1	0.61694102	0.69705461	0.89	0.3789
X1*X1*X1	0.00499808	0.01452264	0.34	0.7317

10.4 Orthogonal and near orthogonal designed experiments

Model choice with orthogonal experiments

Three way interactions between categorical variables follow the same shorthand as between continuous variables in the previous example.

SAS COMMANDS FOR BOX 10.6 **Analysis of a factorial experiment**

```

Commands  proc glm data=gandh.Chapter10;
           class BAC TEMP OXYGEN;
           model ROT = BAC | TEMP | OXYGEN;
           run;

```

Menu route Statistics > Anova > Linear Models...

ROT → Dependent

BAC TEMP OXYGEN → Class

Model

Select 3 as factorial order by clicking on ▲

Independent → Effects in model

(BAC & TEMP & OXYGEN) → Factorial

SAS OUTPUT FOR BOX 10.6 **Analysis of a factorial experiment**

The GLM Procedure

Dependent Variable: ROT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	1863.703704	109.629630	4.68	<.0001
Error	36	843.333333	23.425926		
Corrected Total	53	2707.037037			

R-Square	Coeff Var	Root MSE	ROT Mean
0.688466	51.44918	4.840034	9.407407

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BAC	2	651.8148148	325.9074074	13.91	<.0001
TEMP	1	848.0740741	848.0740741	36.20	<.0001
OXYGEN	2	97.8148148	48.9074074	2.09	0.1387
BAC*OXYGEN	4	30.0740741	7.5185185	0.32	0.8621
BAC*TEMP	2	152.9259259	76.4629630	3.26	0.0498
TEMP*OXYGEN	2	1.5925926	0.7962963	0.03	0.9666
BAC*TEMP*OXYGEN	4	81.4074074	20.3518519	0.87	0.4921

The same commands are required when orthogonality is lost, but the order of the explanatory variables in the model formulae is varied (see main text). Just as in chapter seven, we can use the Model button in menus to determine the order of the model terms in the model formula.

10.5 Looking for trends across levels of a categorical variable

First of all we request two extra pieces of output from the original orthogonal analysis: a table of means for the significant two-way interaction (BAC*TEMP – see Box 10.6), and an interaction diagram. The menu method automatically provides a fancy graph, which invokes a very long set of command behind the scenes. We choose the error bars to be ± 2 standard errors via the 'Plots' button. (You can see all the commands by using the menus, and then inspecting the log file.) In the command language, we create a new dataset using the 'out' option to the 'lsmeans' statement, which contains the least square means in a variable called 'lsmean'. Then we use a shorter set of commands to get a simpler graph without the error bars. We need as many 'SYMBOLi' lines as there are lines to plot, and SAS uses them in turn as each line comes to be drawn. 'CI' is the colour of the interpolating line, and 'LINE=' indicates a degree of dashing, with LINE=1 meaning solid.

SAS COMMANDS FOR BOX 10.9 AND FIGURE 10.4 Means and interaction diagram for ROT by TEMP and BAC

```

Commands  proc glm data=gandh.Chapter10;
           class bac temp oxygen;
           model ROT = BAC | TEMP | OXYGEN;
           lsmeans BAC*TEMP / out=WORK.INDIAG;

           goptions reset=SYMBOL;

           proc goptions;
             SYMBOL1 LINE=1 CI=BLACK INTERPOL=JOIN;
             SYMBOL2 LINE=2 CI=BLUE INTERPOL=JOIN;

           proc gplot data=WORK.INDIAG;
             label lsmean = "Predicted ROT";
             plot lsmean * BAC = TEMP;
           run;

```

Menu route Statistics > Anova > Linear Models...

 ROT → Dependent

 BAC TEMP OXYGEN → Class

 Model

 Select 3 as factorial order by clicking on ▲

 Independent → Effects in model

 (BAC & TEMP & OXYGEN) → Factorial

 Plots with Means tab

Plot dependent means for two-way effects

Predicted means

2 se

 Means with LSMMeans tab

 BAC*TEMP → LS Mean

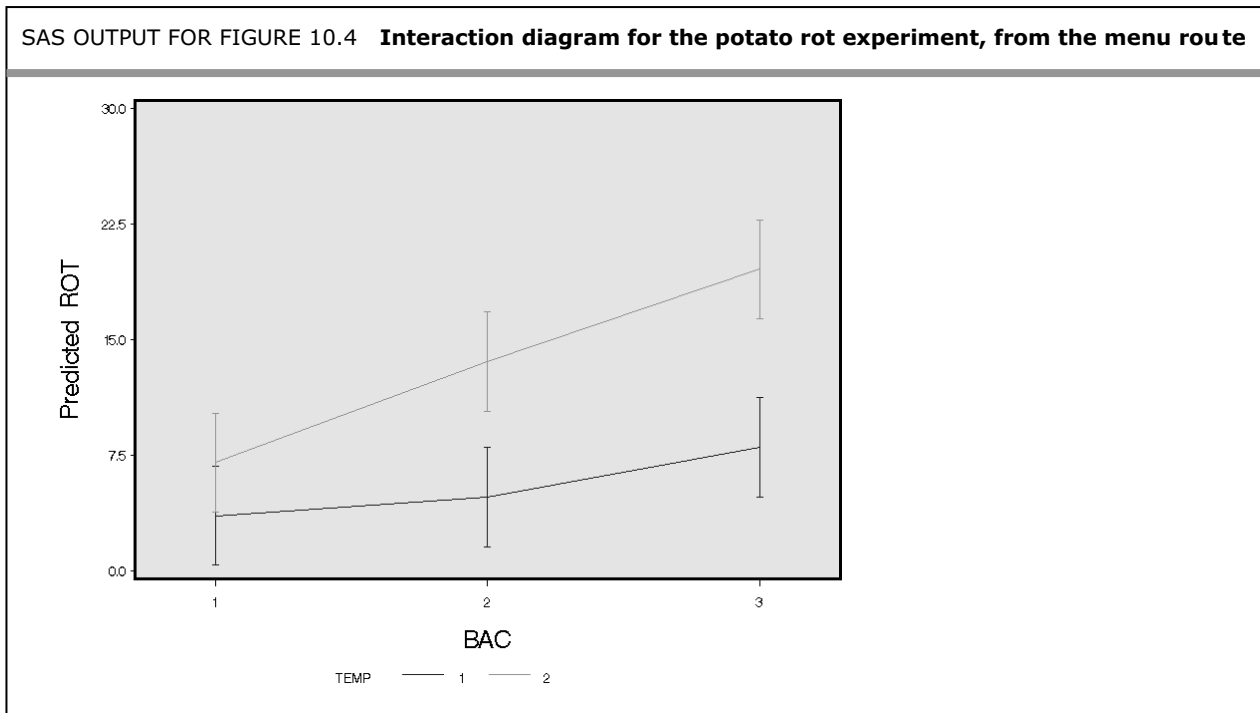
The following extra output would be produced by the LSMEANS statement or 'Plots' and 'Means' menu options:

SAS OUTPUT FOR BOX 10.9 **Degree of rotting in potatoes, according to TEMP and BAC treatments**

The GLM Procedure
Least Squares Means

BAC	TEMP	ROT LSMEAN
1	1	3.5555556
1	2	7.0000000
2	1	4.7777778
2	2	13.5555556
3	1	8.0000000
3	2	19.5555556

The 'Plots' requested in the menu route are all three two-way plots. You need to navigate your way through the results tree to find the relevant two way graph, which is shown below.



Examination of this interaction diagram is what suggests to us that there is a linear relationship between ROT and BAC, and that this may even be curvilinear at TEMP level 1. This then leads on to the analysis in which BAC is declared as continuous and fitted as a polynomial, interacting with TEMP.

SAS COMMANDS FOR BOX 10.10 **Reanalysing the potato rot experiment, looking for trends**

```

Commands  proc glm data=gandh.Chapter10;
           class TEMP OXYGEN;
           model ROT = OXYGEN TEMP | BAC | BAC;
           run;

```

Menu route Statistics > Anova > Linear Models...

ROT → Dependent

TEMP OXYGEN → Class

BAC → Quantitative

Independent → Effects in model

OXYGEN → ADD

(BAC & TEMP) → Factorial

Select BAC in Independent pane, and BAC in 'Effects in model' pane, then press 'Cross'.

Select BAC in Independent pane, and BAC * TEMP in 'Effects in model' pane, then press 'Cross'.

Type I

SAS OUTPUT FOR BOX 10.10 Reanalysing the potato rot experiment, looking for trends					
The GLM Procedure					
Dependent Variable: ROT					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	1750.629630	250.089947	12.03	<.0001
Error	46	956.407407	20.791465		
Corrected Total	53	2707.037037			
	R-Square	Coeff Var	Root MSE	ROT Mean	
	0.646696	48.46995	4.559766	9.407407	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
OXYGEN	2	97.8148148	48.9074074	2.35	0.1065
TEMP	1	848.0740741	848.0740741	40.79	<.0001
BAC	1	650.2500000	650.2500000	31.27	<.0001
BAC*TEMP	1	148.0277778	148.0277778	7.12	0.0105
BAC*BAC	1	1.5648148	1.5648148	0.08	0.7851
BAC*BAC*TEMP	1	4.8981481	4.8981481	0.24	0.6297

10.7 Exercises

Testing polynomials requires sequential sums of squares

SAS COMMANDS FOR BOX 10.11 (A) Analysis based on X	
Commands	<pre>proc glm data=gandh.Chapter10; model Y = X X X; run;</pre>
Menu route	<p>Statistics > Anova > Linear Models...</p> <p>Y → Dependent</p> <p>X → Quantitative</p> <p><input type="text" value="Model"/></p> <p>Select 3 as polynomial order by clicking on ▲</p> <p>Independent → Effects in model</p> <p>X → Polynomial</p> <p><input type="text" value="Statistics"/></p> <p><input checked="" type="checkbox"/> Type 1</p>

SAS COMMANDS FOR BOX 10.11 (A)

Analysis based on X: the first analysis bases the p-values on adjusted sums of squares and the second on sequential sums of squares

The GLM Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	63.75326128	21.25108709	72.70	<.0001
Error	68	19.87586650	0.29229215		
Corrected Total	71	83.62912778			

R-Square	Coeff Var	Root MSE	Y Mean
0.762333	1.683612	0.540641	32.11194

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X	1	58.90626958	58.90626958	201.53	<.0001
X*X	1	4.30534306	4.30534306	14.73	0.0003
X*X*X	1	0.54164864	0.54164864	1.85	0.1779

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X	1	0.71903354	0.71903354	2.46	0.1214
X*X	1	0.15111142	0.15111142	0.52	0.4746
X*X*X	1	0.54164864	0.54164864	1.85	0.1779

The second analysis requires exactly the same commands, substituting XS for X.

Partitioning a sum of squares into polynomial components

The SAS output for this exercise may be found in the answers to exercises.