

10

Model selection I: principles of model choice and designed experiments

10.2 Three principles of model choice

There are two differences between SPSS and the generic language used in the main text that are important for this chapter. First, SPSS's menu route does not allow polynomial terms in a continuous variable. This somewhat complicates the menu approach to many of the analyses in this chapter, as we fool SPSS into doing the required analysis by calculating new variables. The second difference is that there is no shorthand way of writing interaction terms—the model $A B A*B$ in most languages has a shortcut form, such as $A|B$, but not in SPSS. Many design statements are therefore long, and many menu routes laborious, as a consequence. On the other hand, the full factorial model, including all categorical variables, is the default model. Users must not fall into the temptation of fitting the full factorial model just to avoid the labour of spelling out a more appropriate model.

Model choice in the polynomial problem

In the early sections of this chapter in the main text, a polynomial model was built up by fitting the three models X , $X + X^2$ and $X + X^2 + X^3$. Choices can be made between these models on the basis of economy of variables, multiplicity of p-values and considerations of marginality. These models cannot be fitted naturally in SPSS via the menu route (although they can, curiously, via the syntax route). We now demonstrate a technique that manages to fool SPSS into doing the required analyses via menus. At the same time we ask SPSS to test the sequential (Type I) sums of squares. When testing adjusted (Type III) sums of squares, it is necessary to fit three separate models, because only the final p-value is valid. If fitting models for which such considerations of marginality come into play, it is more efficient and informative to request ANOVA tables based on sequential sums of squares, as is shown below for the *simple polynomial* dataset.

To fudge the polynomial terms in the menu route, we calculate the squared and cubed terms in new variables called X1P2 and X1P3 (to represent X1 to the power 2 and X1 to the power 3 respectively), and include them in the model in place of X1*X1 and X1*X1*X1.

SPSS COMMANDS FOR BOX 10.5	
ANOVA tables using sequential rather than adjusted sums of squares	
Syntax	<pre>glm Y1 with X1 /print parameters /method sstype(1) /design X1 X1*X1 X1*X1*X1.</pre>
Menu route	<p>Window > Data Editor</p> <p>Transform > Compute</p> <p style="padding-left: 40px;">X1P2 → Target Variable</p> <p style="padding-left: 40px;">X1*X1 → Numeric Expression</p> <p>Transform > Compute</p> <p style="padding-left: 40px;">X2P3 → Target Variable</p> <p style="padding-left: 40px;">X1*X1*X1 → Numeric Expression</p> <p>Analyze > General Linear Model > Univariate</p> <p style="padding-left: 40px;">Y1 → Dependent Variable</p> <p style="padding-left: 40px;">X1 X1P2 X1P3 → Covariate(s)</p> <p style="padding-left: 40px;"><input type="button" value="Options"/></p> <p style="padding-left: 80px;"><input checked="" type="checkbox"/> Parameter estimates</p> <p style="padding-left: 40px;"><input type="button" value="Model"/></p> <p style="padding-left: 80px;">Sum of Squares: <input type="button" value="Type I ▼"/></p>

SPSS OUTPUT FOR BOX 10.5

ANOVA tables using sequential rather than adjusted sums of squares**General linear model****Tests of Between-Subjects Effects**

Dependent Variable: Y1

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6919889.2 ^a	3	2306629.729	379.382	.000
Intercept	7513185.6	1	7513185.620	1235.727	.000
X1	6663020.6	1	6663020.646	1095.897	.000
X1 * X1	256148.401	1	256148.401	42.130	.000
X1 * X1 * X1	720.139	1	720.139	.118	.732
Error	462077.908	76	6079.972		
Total	14895153	80			
Corrected Total	7381967.1	79			

a. R Squared = .937 (Adjusted R Squared = .935)

Parameter Estimates

Dependent Variable: Y1

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	-15.748	33.924	-.464	.644	-83.315	51.818
X1	6.179	9.625	.642	.523	-12.991	25.349
X1 * X1	.617	.697	.885	.379	-.771	2.005
X1 * X1 * X1	4.998E-03	.015	.344	.732	-2.393E-02	3.392E-02

10.4 Orthogonal and near orthogonal designed experiments**Model choice with orthogonal experiments**

The DESIGN subcommand and the MODEL subdialog are used here to make clear the nature of the model being fitted. They are long-winded, however, and in this case can both be omitted altogether. This is because the design is the SPSS's default design, which contains the full factorial combination of all categorical variables, plus (though in this case there aren't any) the continuous variables without any of their interactions. We repeat the analyses, once with adjusted and once with sequential sums of squares. The commands in bold will produce the sequential sums of squares, and omitting these commands will produce the adjusted sums of squares (although this also assumes you press the 'reset' button in the main 'univariate' window between analyses in the menu route). Remember the convention that BAC & TEMP should be interpreted as selecting the two variables together, before clicking on the 'build terms' arrow.

SPSS COMMANDS FOR BOX 10.6 **Analysis of a factorial experiment.**

Perform once with and once without the bold lines (pressing reset between analyses in the menu route)

Syntax `glm ROT by BAC TEMP OXYGEN`
 `/method sstype(1)`
 `/design BAC TEMP OXYGEN BAC*TEMP BAC*OXYGEN`
 `TEMP*OXYGEN BAC*TEMP*OXYGEN.`

Menu route Analyze > General Linear Model > Univariate

 ROT → Dependent Variable

 BAC TEMP OXYGEN → Fixed Factor(s)

 Model

Custom

 Factors & Covariates → Build Terms → Model

 BAC → BAC

 TEMP → TEMP

 OXYGEN → OXYGEN

 BAC & TEMP → BAC*TEMP

 BAC & OXYGEN → BAC*OXYGEN

 TEMP & OXYGEN → TEMP*OXYGEN

 BAC & TEMP & OXYGEN → BAC*TEMP*OXYGEN

 Sum of Squares: Type I ▼

SPSS OUTPUT FOR BOX 10.6 **Analysis of a factorial experiment****General linear model****Between-Subjects Factors**

		N
BAC	1	18
	2	18
	3	18
TEMP	1	27
	2	27
OXYGEN	1	18
	2	18
	3	18

Tests of Between-Subjects Effects

Dependent Variable: ROT

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1863.704 ^a	17	109.630	4.680	.000
Intercept	4778.963	1	4778.963	204.003	.000
BAC	651.815	2	325.907	13.912	.000
TEMP	848.074	1	848.074	36.202	.000
OXYGEN	97.815	2	48.907	2.088	.139
BAC * TEMP	152.926	2	76.463	3.264	.050
BAC * OXYGEN	30.074	4	7.519	.321	.862
TEMP * OXYGEN	1.593	2	.796	.034	.967
BAC * TEMP * OXYGEN	81.407	4	20.352	.869	.492
Error	843.333	36	23.426		
Total	7486.000	54			
Corrected Total	2707.037	53			

a. R Squared = .688 (Adjusted R Squared = .541)

*and the second analysis produces this ANOVA table instead....***Tests of Between-Subjects Effects**

Dependent Variable: ROT

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1863.704 ^a	17	109.630	4.680	.000
Intercept	4778.963	1	4778.963	204.003	.000
BAC	651.815	2	325.907	13.912	.000
TEMP	848.074	1	848.074	36.202	.000
OXYGEN	97.815	2	48.907	2.088	.139
BAC * TEMP	152.926	2	76.463	3.264	.050
BAC * OXYGEN	30.074	4	7.519	.321	.862
TEMP * OXYGEN	1.593	2	.796	.034	.967
BAC * TEMP * OXYGEN	81.407	4	20.352	.869	.492
Error	843.333	36	23.426		
Total	7486.000	54			
Corrected Total	2707.037	53			

a. R Squared = .688 (Adjusted R Squared = .541)

The following output would then be produced in addition to the ANOVA table:

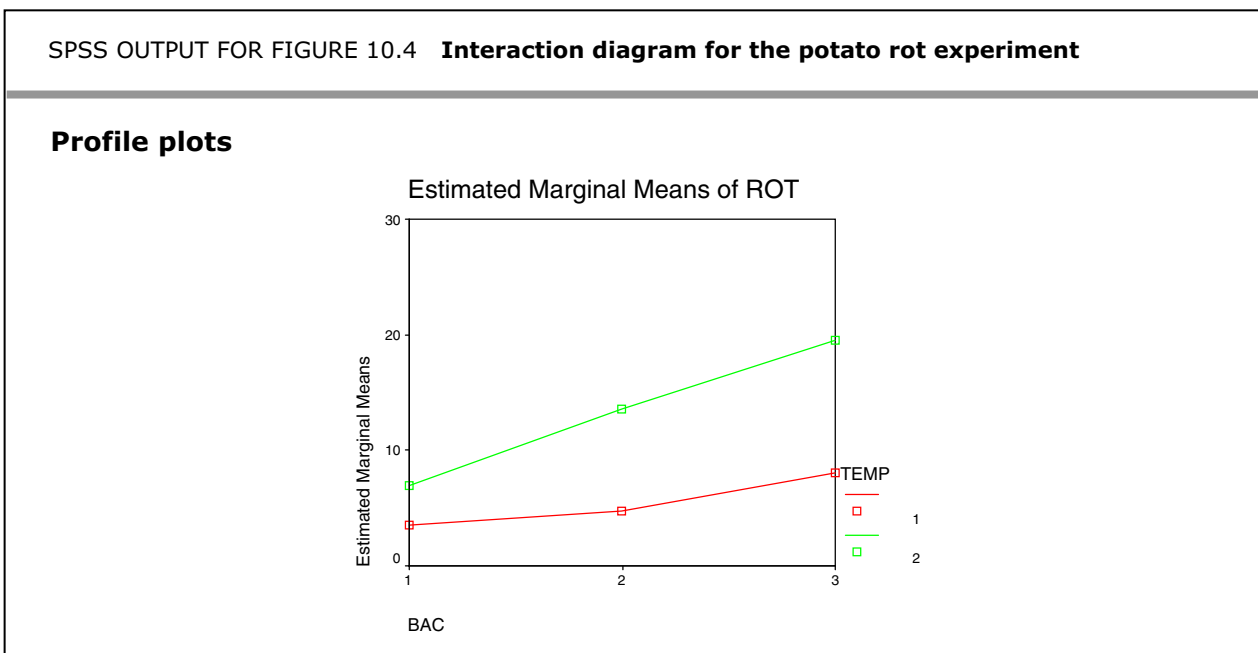
SPSS OUTPUT FOR BOX 10.9
Degree of rotting in potatoes, according to TEMP and BAC treatments

Estimated marginal means

BAC * TEMP

Dependent Variable: ROT

BAC	TEMP	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	3.556	1.613	.284	6.828
	2	7.000	1.613	3.728	10.272
2	1	4.778	1.613	1.506	8.050
	2	13.556	1.613	10.284	16.828
3	1	8.000	1.613	4.728	11.272
	2	19.556	1.613	16.284	22.828



Examination of this interaction diagram is what suggests to us that there is a directional 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. The output for the syntax route and menu route are given separately. Notice that the calculated variable BAC2 plays exactly the same role in the menu output as the quadratic term BAC*BAC does in the more conventional syntax route.

SPSS COMMANDS FOR BOX 10.10

Reanalysing the potato rot experiment, looking for trends

Syntax `glm ROT by TEMP OXYGEN with BAC`
 `/method sstype(1)`
 `/design OXYGEN TEMP BAC BAC*TEMP BAC*BAC`
 `BAC*BAC*TEMP.`

Menu route **Window > Data Editor**
 Transform > Compute
 BAC2 → Target Variable
 BAC*BAC → Numeric Expression
 Analyze > General Linear Model > Univariate
 ROT → Dependent Variable
 TEMP OXYGEN → Fixed Factor(s)
 BAC BAC2 → Covariate(s)
 Model
 Custom
 Factors & Covariates → Build Terms → Model
 OXYGEN → OXYGEN
 TEMP → TEMP
 BAC → BAC
 BAC & TEMP → BAC*TEMP
 BAC2 → BAC2
 BAC2 & TEMP → BAC2*TEMP
 Sum of Squares: Type I ▼

SPSS OUTPUT FOR BOX 10.10 (SYNTAX ROUTE)

Reanalysing the potato rot experiment, looking for trends**General linear model****Between-Subjects Factors**

		N
TEMP	1	27
	2	27
OXYGEN	1	18
	2	18
	3	18

Tests of Between-Subjects Effects

Dependent Variable: ROT

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1750.630 ^a	7	250.090	12.028	.000
Intercept	4778.963	1	4778.963	229.852	.000
OXYGEN	97.815	2	48.907	2.352	.106
TEMP	848.074	1	848.074	40.790	.000
BAC	650.250	1	650.250	31.275	.000
TEMP * BAC	148.028	1	148.028	7.120	.010
BAC * BAC	1.565	1	1.565	.075	.785
TEMP * BAC * BAC	4.898	1	4.898	.236	.630
Error	956.407	46	20.791		
Total	7486.000	54			
Corrected Total	2707.037	53			

a. R Squared = .647 (Adjusted R Squared = .593)

SPSS OUTPUT FOR BOX 10.10 (MENU ROUTE)

Reanalysing the potato rot experiment, looking for trends**Univariate analysis of variance****Between-Subjects Factors**

		N
TEMP	1	27
	2	27
OXYGEN	1	18
	2	18
	3	18

Tests of Between-Subjects Effects

Dependent Variable: ROT

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1750.630 ^a	7	250.090	12.028	.000
Intercept	4778.963	1	4778.963	229.852	.000
OXYGEN	97.815	2	48.907	2.352	.106
TEMP	848.074	1	848.074	40.790	.000
BAC	650.250	1	650.250	31.275	.000
TEMP * BAC	148.028	1	148.028	7.120	.010
BAC ²	1.565	1	1.565	.075	.785
TEMP * BAC ²	4.898	1	4.898	.236	.630
Error	956.407	46	20.791		
Total	7486.000	54			
Corrected Total	2707.037	53			

a. R Squared = .647 (Adjusted R Squared = .593)


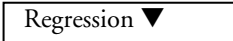
We now plot an interaction plot containing both the fitted lines and data. In both routes, we ask for a separate line for each group. The groups are defined by using the variable TEMP in a legend, in this case separating the groups by colour. The lines will not be drawn, however, if TEMP is defined as having a scale of measurement, so we have to change this. In the syntax route, we just declare TEMP as being CATEGORICAL. In the menu route, we need to make this change in the data window. One technicality that is different here is that the variables are moved to their appropriate boxes by selecting them, and then dragging with the mouse (rather than clicking on a particular arrow to effect the transfer). This is indicated by (drag and drop) between two arrows.

SPSS COMMANDS FOR BOX 10.6 Presenting the results for rotting potatoes

Syntax igraph
 /viewname='scatterplot'
 /x1 = var(BAC) type = scale
 /y = var(ROT) type = scale
 /color = var(TEMP) type = categorical
 /coordinate = vertical
 /fitline method = regression linear line = meffect
 /scatter.

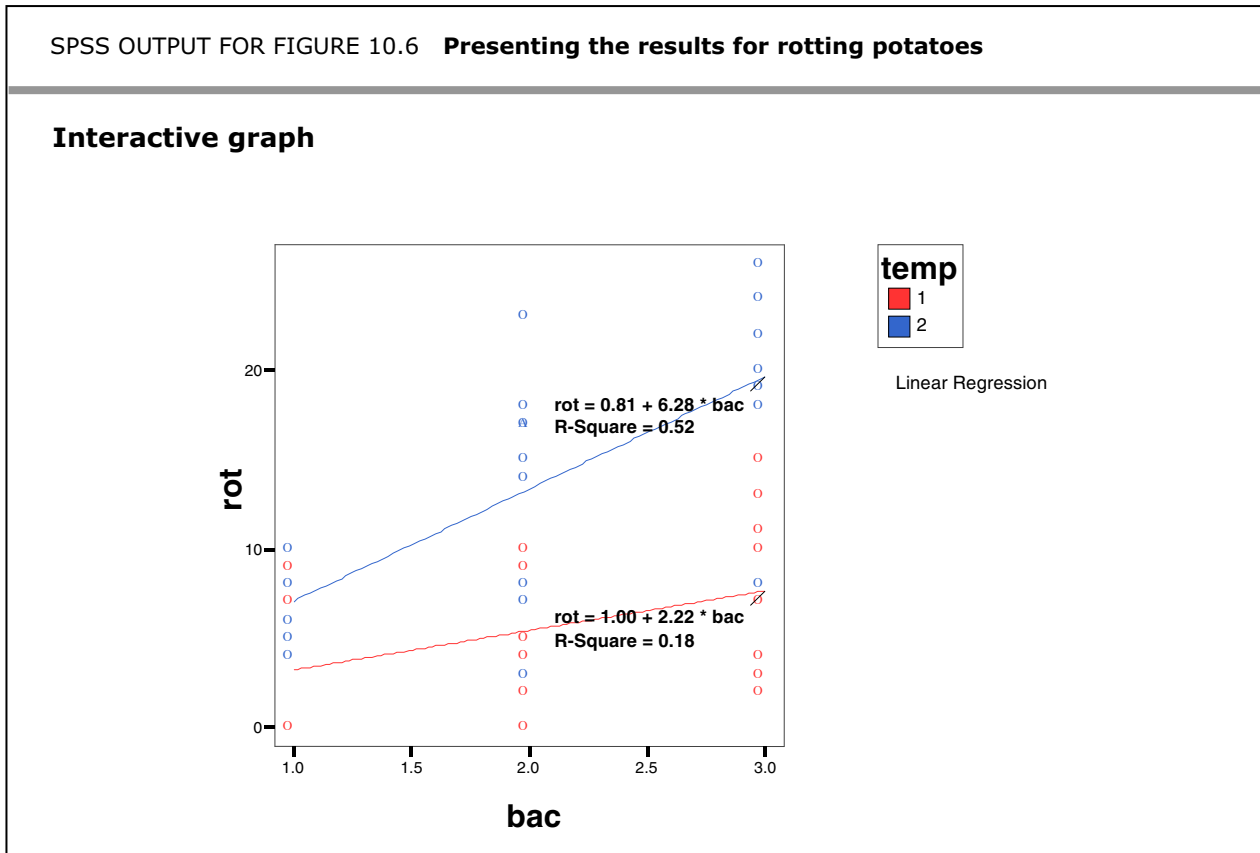
Menu route Window > SPSS Data Editor
 Select 'VariableView' among tabs at the bottom left
 Click on the cell in the 'TEMP' row and 'Measurement' column
 From the pull-down menu select 'Ordinal'. You may need to re-import
 the chapter10data.sav file for this change to take effect.

 Graphs > Interactive > Scatterplot
 ROT → (drag and drop) → (Y Axis box)
 BAC → (drag and drop) → (X Axis box)
 TEMP → (drag and drop) → Color


 Method: 

 Fit lines for:
 – Total
 Subgroups

This would produce the following graph (you may not see the colour below depending on how you are viewing the file):



10.7 Exercises

Testing polynomials requires sequential sums of squares

SPSS COMMANDS FOR BOX 10.11 (a) **Analysis based on X.**

Perform once without and once with the bold lines (pressing reset between analyses in the menu route)

Syntax	<pre>glm Y with X /method sstype(1) /design X X*X X*X*X.</pre>
--------	---

Menu route	<p>Window > Data Editor</p> <p>Transform > Compute</p> <p style="padding-left: 40px;">X2 → Target Variable</p> <p style="padding-left: 40px;">X*X → Numeric Expression</p> <p>Transform > Compute</p> <p style="padding-left: 40px;">X3 → Target Variable</p> <p style="padding-left: 40px;">X*X*X → Numeric Expression</p> <p>Analyze > General Linear Model > Univariate</p> <p style="padding-left: 40px;">Y → Dependent Variable</p> <p style="padding-left: 40px;">X X2 X3 → Covariate(s)</p> <div style="border: 1px solid black; width: fit-content; padding: 2px; margin-left: 80px;">Model</div> <p style="padding-left: 80px;">Sum of Squares: Type I ▼</p>
------------	--

SPSS OUTPUT FOR BOX 10.11 (a) **Analysis based on X**

General linear model

Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	63.753 ^a	3	21.251	72.705	.000
Intercept	3256.398	1	3256.398	11140.902	.000
X	.719	1	.719	2.460	.121
X * X	.151	1	.151	.517	.475
X * X * X	.542	1	.542	1.853	.178
Error	19.876	68	.292		
Total	74328.371	72			
Corrected Total	83.629	71			

a. R Squared = .762 (Adjusted R Squared = .752)

and the second analysis produces the following ANOVA table instead:

Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	63.753 ^a	3	21.251	72.705	.000
Intercept	74244.742	1	74244.742	254008.7	.000
X	58.906	1	58.906	201.532	.000
X * X	4.305	1	4.305	14.730	.000
X * X * X	.542	1	.542	1.853	.178
Error	19.876	68	.292		
Total	74328.371	72			
Corrected Total	83.629	71			

a. R Squared = .762 (Adjusted R Squared = .752)

The second analysis requires exactly the same commands, substituting XS for X, and in the menu route calculating XS2 and XS3 from XS.

Partitioning a sum of squares into polynomial components

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