

3

Models, parameters and GLMS

In SPSS, the general linear model command (GLM) can be used to analyse models with both continuous and categorical explanatory variables, instead of using the REGRESSION command for continuous variables. This chapter will briefly summarise the relevant conventions. For example, you introduce SPSS to categorical variables using the keyword BY, and to continuous variables using the keyword WITH.

3.2 Expressing all models as linear equations

The commands to conduct a general linear model analysis on the *trees* data set are presented below. The ‘design’ subcommand gives the right hand side of the model formulae used in the main text (the y-variable is named in the ‘glm’ line itself). In this case, we could do without the design subcommand, because SPSS makes up a default model from the x-variables named in the glm line, and in this case the default model is the one we want. In future, we will always use the design subcommand even when we do want the default model, because focussing on the exact model formula is such an important part of the approach taken in the main text. The “PRINT PARAMETER” subcommand and the check box for parameter estimates ensure that SPSS provides the parameter estimates.

SPSS COMMANDS FOR BOX 3.1 General linear models with continuous explanatory variables	
Syntax	<pre>glm VOLUME with HEIGHT /print parameter /design HEIGHT.</pre>
Menu route	Analyze > General Linear Model > Univariate VOLUME → Dependent Variable HEIGHT → Covariate(s) <input type="text" value="Options"/> <input checked="" type="checkbox"/> Parameter estimates

Note that we introduced HEIGHT using the WITH keyword, as it is a continuous variable. This will produce the following output:

SPSS OUTPUT FOR BOX 3.1 General linear model for the trees data set						
General linear model						
Tests of Between-Subjects Effects						
Dependent Variable: VOLUME						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	
Corrected Model	2901.189 ^a	1	2901.189	16.164	.000	
Intercept	1589.819	1	1589.819	8.858	.006	
HEIGHT	2901.189	1	2901.189	16.164	.000	
Error	5204.895	29	179.479			
Total	36324.990	31				
Corrected Total	8106.084	30				
a. R Squared = .358 (Adjusted R Squared = .336)						
Parameter Estimates						
Dependent Variable: VOLUME						
Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	-87.124	29.273	-2.976	.006	-146.994	-27.253
HEIGHT	1.543	.384	4.021	.000	.758	2.328

It can be seen that exactly the same information is given in the output. All those items that are missing (the fitted equation and s for example) can all be calculated from the 'Tests of Between-Subjects Effects' table and the 'Parameter Estimates' table, as discussed in the main text. These are SPSS's names for the ANOVA table and coefficients table, respectively. The ANOVA table tests *whether* the two variables are related. (Does tree height give us any information about tree volume? — The answer is yes, with a p-value of less than 0.0005). The coefficient table tells us *how* the variables are related ($VOLUME = -87.124 + 1.543 \text{ HEIGHT}$). SPSS automatically provides Type III sums of squares in the ANOVA table, which are equivalent to the adjusted sums of squares in the main text. The meanings of sequential and adjusted sums of squares, and when it is appropriate to use one or the other, are discussed in detail in the main text.

The form of the ANOVA table in SPSS disguises some important facts, and also uses terms differently from most statistical packages, and from the main text. Here we make a comparison with the terms used in the main text.

1. In the source column, there is both 'total' and 'corrected total'. SPSS Total has no equivalent in the main text, but is simply the sum of the squares of the values of the dependent variable (VOLUME in this case). The corresponding degrees of freedom are therefore the total number of datapoints.
2. SPSS Corrected Total = Total sum of squares in the main text = HEIGHT + Error. Therefore the degrees of freedom associated with the corrected total correspond to the total degrees of freedom in the main text (n-1).
3. Total DF = Corrected Total DF + Intercept DF. This corresponds to the first step in an ANOVA, namely fitting the grand mean and calculating the total sum of squares. The difference between SST and the sum of squares of the y values can be thought of as the intercept sum of squares, or the sum of squares explained by fitting the grand mean.
4. Corrected Model Sum of squares and DF = the model or regression sum of squares and degrees of freedom. This is equivalent to the sum of the squares explained by all explanatory variables or the sum of terms between 'intercept' and 'error', not including either (in this case, just HEIGHT).

The terms that add up are not adjacent (e.g. HEIGHT and Error add up to Corrected Total), and the whole table seems designed to obscure these relationships. On the other hand, the extra division (Total = Corrected Total + Intercept) does illustrate the partitioning of SS and DF for the first step of ANOVA, which is often excluded from ANOVA tables. It also allows us to test whether the intercept (or the grand mean) is significantly different from zero in the ANOVA table as well as the coefficients table. . If we had requested the Type I or sequential sums of squares, then the SS would have added up in the same pattern as the DF for all the relationships. As it is, relationship 2 is the same as shown in the main text, and the sums of squares do add up for that particular relationship.

One final point about the output is that the menu route produces a different heading (it says Univariate Analysis of Variance) from the syntax route (General Linear Model). This is nothing to worry about.

Similarly, the commands for analysing the *fertiliser* data set are given below: Note how on this occasion we use the BY keyword as it is a categorical variable.

SPSS COMMANDS FOR BOX 3.2 General linear models with categorical explanatory variables	
Syntax	<pre>glm YIELD by FERTIL /print parameter /design FERTIL.</pre>
Menu route	Analyze > General Linear Model > Univariate YIELD → Dependent Variable FERTIL → Fixed Factor(s) <input type="checkbox"/> Options <input checked="" type="checkbox"/> Parameter estimates

This will give the following output:

SPSS OUTPUT FOR BOX 3.2 General linear model for the <i>fertiliser</i> dataset						
General linear model						
Between-Subjects Factors						
		N				
FERTIL	1	10				
	2	10				
	3	10				
Tests of Between-Subjects Effects						
Dependent Variable: YIELD						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	
Corrected Model	10.823 ^a	2	5.411	5.702	.009	
Intercept	646.909	1	646.909	681.697	.000	
FERTIL	10.823	2	5.411	5.702	.009	
Error	25.622	27	.949			
Total	683.354	30				
Corrected Total	36.445	29				
a. R Squared = .297 (Adjusted R Squared = .245)						
Parameter Estimates						
Dependent Variable: YIELD						
Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4.487	.308	14.566	.000	3.855	5.119
[FERTIL=1]	.958	.436	2.199	.037	6.411E-02	1.852
[FERTIL=2]	-.488	.436	-1.120	.273	-1.382	.406
[FERTIL=3]	0 ^a
a. This parameter is set to zero because it is redundant						

The frightening warning at the bottom of the final table of output reflects a simple statistical reality, that with three levels of FERTIL, there are only two differences between them. This means that two parameters can be estimated, not three, but there is no unique way of choosing which two parameters. SPSS's choice is always to set the parameter for the final level to zero, as indicated in the output. Because it is bound to be zero, the standard error is also zero (not shown in the output), and there is no sense in testing whether it does equal zero. This method of so-called 'aliasing' is not the same as that used in the main text. However, all important statistical results are identical. Aliasing affects the presentation of the model, not its reality.

To reconstruct the group means from the coefficients table, take the intercept of 4.487 and add the deviation for fertiliser 1, 2 or 3. As the deviation for fertiliser 3 has been set to zero, the intercept is also the estimate for fertiliser 3. This is also described in the main text, with reference to Table 3.3.

With the SPSS general linear modelling facility, it is also possible to request a table of the means, as shown below:

SPSS COMMANDS FOR TABLE 3.3 Obtaining least squares means for a categorical variable	
Syntax	<pre>glm YIELD by FERTIL /emmeans = tables(FERTIL) /print parameter /design FERTIL.</pre>
Menu route	<p>Analyze > General Linear Model > Univariate</p> <p>YIELD → Dependent Variable</p> <p>FERTIL → Fixed Factor(s)</p> <p>Options</p> <p><input checked="" type="checkbox"/> Parameter estimates</p> <p>FERTIL → Display Means for</p>

This will produce the table below in addition to the output shown in Box 3.2. The command EMMEANS stands for 'estimated marginal means', and is a synonym for the 'least squares means' and 'adjusted means' of other packages. In this simple case, it is just the mean YIELD in the different FERTIL groups.

SPSS OUTPUT FOR TABLE 3.3 **Obtaining the group means for the fertiliser dataset****Estimated marginal means****FERTIL**

Dependent Variable: YIELD

FERTIL	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	5.445	.308	4.813	6.077
2	3.999	.308	3.367	4.631
3	4.487	.308	3.855	5.119

3.3 Turning the tables and creating datasets

SPSS has a powerful language in which it is possible to perform statistical simulations. We will not use the pure menu route, or the pure syntax route, as each is very masochistic in its own way. Instead, we use a convenient combination.

The chapter 3 dataset has YIELD and FERTIL from the fertiliser dataset. We will make use of the variable FERTIL, but also define additional columns. The parameters in the simulation are K3, K1, K2 and SIGMA. These are represented here as full length variables, but we will assign values so that the value of K3, for example, is the same for every datapoint—otherwise we would be allowing different datapoints to have different intercepts. Here is the syntax file, available as Chapter3sim.sps on the website, or you may prefer to type it in.

SPSS COMMANDS FOR BOX 3.3a Creating your own data set	
<pre>compute DUM1=(FERTIL=1) . compute DUM2=(FERTIL=2) . execute .</pre>	<i>First group</i>
<pre>compute K3=10.7 . compute K1=3.4 . compute K2=1.1 . compute SIGMA=2.3 . execute.</pre>	<i>Second group</i>
<pre>compute NOISE=rv.normal(0,1) . compute Y= K3 + K1*DUM1 + K2*DUM2 + SIGMA*NOISE . execute. glm Y by FERTIL /print parameter /design FERTIL .</pre>	<i>Third group</i>

The first group of commands sets up the dummy variables. SPSS's dummy variables are not the same as in the main text, reflecting the different method of aliasing. Instead DUM1 has a 1 wherever FERTIL=1, and a 0 elsewhere, and DUM2 has a 1 wherever FERTIL=2, and a 0 elsewhere. It is important to emphasise again that this difference does not affect any of the statistical conclusions reached. The consequence of this choice is that the constant, K3, is the mean we expect for FERTIL=3. The mean we expect for FERTIL=1 is K3+K1 and the mean we expect for FERTIL=2 is K3+K2. These are different rules to those that apply in the main text, because SPSS uses a different form of 'aliasing'. Just as in the main text, however, if you make SIGMA small enough, you will be able to confirm that these formulae for the means of groups are indeed correct.

You use them as follows: First make sure have the dataset Chapter3data.sav open, (or import the file Chapter3data.xls), and then you need to find the syntax window to type commands in. If you open the file Chapter3sim.sps, it will open as a syntax file automatically, and you're fine. If you want to type the commands in, choose Window > Syntax to get the syntax window. If there is no Syntax window available, then choose File > New > Syntax to create one. You're ready to go.

Select the first group of commands by dragging over them with the mouse, including the final EXECUTE with its period. Then choose Run > Selection. You should notice in the dataset that DUM1 and DUM2 have their values filled in. Then edit the second group to choose whichever parameter values you want to simulate; then select the second group, and again choose Run > Selection. Four new columns, K3, K1, K2 and SIGMA will appear.

34 Models, parameters and GLMS

Finally, select the third group and choose Run > Selection. This puts standard Normal random variables into 'NOISE', and creates 'Y' according to the model, and also analyses it with GLM. The output window will come to the front and you will see the analysis of variance table, an example of which is shown in the box below. Record the information you want, if necessary scrolling down to the parameter estimates. Then choose Window > Syntax, and then Run > Selection, to perform another simulation with the same parameter values but fresh random variables in 'noise'. Record the information you want. Repeat as often as you like!

SPSS OUTPUT FOR BOX 3.4 **Recovering the parameter estimates**

General linear model

Between-Subjects Factors

	N
FERTIL 1	10
2	10
3	10

Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	27.443 ^a	2	13.721	2.652	.089
Intercept	4143.015	1	4143.015	800.710	.000
FERTIL	27.443	2	13.721	2.652	.089
Error	139.703	27	5.174		
Total	4310.161	30			
Corrected Total	167.146	29			

a. R Squared = .164 (Adjusted R Squared = .102)

Parameter Estimates

Dependent Variable: Y

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	10.910	.719	15.168	.000	9.435	12.386
[FERTIL=1]	2.179	1.017	2.142	.041	9.177E-02	4.266
[FERTIL=2]	.344	1.017	.338	.738	-1.743	2.432
[FERTIL=3]	0 ^a

a. This parameter is set to zero because it is redundant

To change the parameters, all you need do is edit the second group to the values you want, select it and choose Run > Selection. Then again select the third group and repeat the Run > Selection, record, Window > Syntax cycle as desired.

The means for the three levels of F can be calculated from the coefficients table. As mentioned earlier, SPSS uses a different form of aliasing, using the mean of group 3 as the reference point. If the intercept is denoted μ , and the parameters given for FERTIL level 1 and 2 as α_1 and α_2 respectively, then the group means may be calculated from the parameters given in the SPSS parameter estimates table as follows:

SPSS VERSION OF TABLE 3.2	
Fertiliser	Mean
1	$\mu + \alpha_1$
2	$\mu + \alpha_2$
3	μ

Here is a hint if you want to extend your simulating. A good way to find out what other facilities are available within COMPUTE is to use the menu route, where they are all laid out. If you have a syntax window open, you can use the menu route to paste new instructions into it. This is done as follows: when the Data window is uppermost, choose Transform > Compute.... then explore the options and choose the command you want, transferring it to the 'Numeric Expression' pane. You need to type the name of a new 'target variable' in the top left hand box. When its a valid instruction, the 'Paste' button at the bottom of the 'Compute Variable' window will be active. Press it, and although your instruction will not be executed, the syntax instructions to execute it will have been pasted into the syntax window. Now you can see how to issue that instruction directly in the syntax window, and it is easy to edit it there.

3.5 Exercises

How variability in the population will influence our analysis

See SPSS output for this exercise in the answers for exercises.