

6

Combining continuous and categorical variables

6.2 Combining continuous and categorical variables

Looking for a treatment for leprosy

To combine continuous and categorical variables is straightforward. We use both keywords BY and WITH in the syntax, or we add variables to both the Fixed Factor(s) and Covariate(s) panes in the menus. This is illustrated below with the *leprosy* dataset.

SPSS COMMANDS FOR BOX 6.1 A treatment for leprosy

Syntax `glm BACAFTER with BACBEF by TREATMT`
 `/print parameters`
 `/design BACBEF TREATMT.`

Menu route Analyze > General Linear Model > Univariate
 BACAFTER → Dependent Variable
 BACBEF → Covariate(s)
 TREATMT → Fixed Factor(s)

Custom

Factors & Covariates → Build Terms → Model

BACBEF	→	BACBEF
TREATMT	→	TREATMT

Parameter estimates

SPSS OUTPUT FOR BOX 6.1 **A treatment for leprosy****General linear model****Between-Subjects Factors**

	N
TREATMT 1	10
2	10
3	10

Tests of Between-Subjects Effects

Dependent Variable: BACAFTER

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	670.827 ^a	3	223.609	15.293	.000
Intercept	7.097E-04	1	7.097E-04	.000	.994
BACBEF	515.015	1	515.015	35.224	.000
TREATMT	83.347	2	41.674	2.850	.076
Error	380.152	26	14.621		
Total	3975.441	30			
Corrected Total	1050.979	29			

a. R Squared = .638 (Adjusted R Squared = .597)

Parameter Estimates

Dependent Variable: BACAFTER

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	2.303	2.103	1.095	.283	-2.019	6.626
BACBEF	.883	.149	5.935	.000	.577	1.189
[TREATMT=1]	-3.906	1.733	-2.254	.033	-7.467	-.344
[TREATMT=2]	-3.042	1.714	-1.775	.088	-6.565	.481
[TREATMT=3]	0 ^a

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

Once again the last treatment (level 3) has been aliased in the table of parameter estimates. The coefficients for the other two treatments represent the mean deviations of those treatments from treatment 3. The coefficients may be used to calculate the fitted values as follows:

$$\text{BACAFTE}R = 2.303 + 0.8831 \times \text{BACBEF} + \begin{pmatrix} \text{TREATMT} \\ 1 & -3.906 \\ 2 & -3.042 \\ 3 & 0 \end{pmatrix}$$

This would give the following three equations:

$$\begin{aligned} \text{BACAFTE}R &= -1.603 + 0.8831 \times \text{BACBEF} && \text{TREATMT } 1 \\ \text{BACAFTE}R &= -0.739 + 0.8831 \times \text{BACBEF} && \text{TREATMT } 2 \\ \text{BACAFTE}R &= 2.303 + 0.8831 \times \text{BACBEF} && \text{TREATMT } 3 \end{aligned}$$

which correspond to the equations on page 99 of the main text.

Sex differences in the weight-fat relationship

With the *fats* dataset, combining continuous and categorical variables can reveal hitherto hidden relationships, as discussed in the main text. To tease apart these hidden relationships it is often useful to specify the order of variables in the model formula. In the syntax route, we write the model formula directly using the DESIGN subcommand, and so we automatically specify the order of terms in the model. In the menu route we can specify a particular order using the model button, and specifying our custom built model. The order in which we place variables in the Model pane then becomes the order in which the variables are fitted.

The key point here is that we wish to fit WEIGHT as the first term in the model, and SEX as the second. In this exercise, we rely on the adjusted, or type III, sums of squares for our F ratio tests, so the order doesn't actually make any difference. However, the order will be crucial when we use sequential, or type I, sums of squares, as we do later in this chapter. You could explore this for yourself by fitting the variables in different orders using sequential sums of squares, and comparing your output with Box 6.2 of the main text.

SPSS COMMANDS FOR BOX 6.2 **Both weight and sex as explanatory variables for fat content**

Syntax `glm FAT with WEIGHT by SEX`
 `/print parameters`
 `/design WEIGHT SEX.`

Menu route `Analyze > General Linear Model > Univariate`
 `FAT → Dependent Variable`
 `WEIGHT → Covariate(s)`
 `SEX → Fixed Factor(s)`

 Custom
 `Factors & Covariates → Build Terms → Model`
 `WEIGHT → WEIGHT`
 `SEX → SEX`

 Parameter estimates

SPSS OUTPUT FOR BOX 6.2 **Both weight and sex as explanatory variables for fat content****General linear model****Between-Subjects Factors**

		N
SEX	1	9
	2	10

Tests of Between-Subjects Effects

Dependent Variable: FAT

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	177.426 ^a	2	88.713	34.624	.000
Intercept	60.425	1	60.425	23.583	.000
WEIGHT	87.105	1	87.105	33.996	.000
SEX	176.098	1	176.098	68.729	.000
Error	40.995	16	2.562		
Total	15509.000	19			
Corrected Total	218.421	18			

a. R Squared = .812 (Adjusted R Squared = .789)

Parameter Estimates

Dependent Variable: FAT

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	9.059	3.000	3.019	.008	2.699	15.418
WEIGHT	.217	.037	5.831	.000	.138	.296
[SEX=1]	7.904	.953	8.290	.000	5.883	9.925
[SEX=2]	0 ^a

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

6.3 Orthogonality in the context of continuous and categorical variables

In many of the subsequent examples, we need to execute the analysis twice — once to obtain the adjusted sums of squares (or Type III, produced by SPSS by default) and once to produce the sequential or Type I sums of squares. In the previous chapters we have illustrated this very explicitly by repeating the commands or menus, with the one additional line. Here we introduce a new convention in which the additional part is in bold. You will need to execute these commands twice — once with and once without the bold part. If you are using the menu route, the sum of squares you selected for your last analysis will remain until you change it, close the programme down at the end of the session, or you can return to the default settings by clicking on the reset button in the main ‘Univariate’ pane. Via the syntax route, however, the default will always be Type III, or adjusted, sums of squares unless you specify otherwise.

This example illustrates orthogonality between a continuous and categorical explanatory variable using the *bacterial growth* dataset.

SPSS COMMANDS FOR BOX 6.3 **Bacterial growth at two levels of lactose.**
Execute once without and once with the bold commands

Syntax `glm BACTERIA by LACTOSE with DAY`
 `/print parameters`
 `/method sstype(1)`
 `/design DAY LACTOSE.`

Menu route Analyze > General Linear Model > Univariate
 BACTERIA → Dependent Variable
 DAY → Covariate(s)
 LACTOSE → Fixed Factor(s)
 Model
 ⊙ Custom
 Factors & Covariates → Build Terms → Model
 DAY → DAY
 LACTOSE → LACTOSE
 Sum of Squares: Type I ▼
 Options
 Parameter estimates

SPSS OUTPUT FOR BOX 6.3 WITH BOTH TYPES OF SUMS OF SQUARE
Bacterial growth at two levels of lactose

General linear model

Between-Subjects Factors

		N
LACTOSE	1	10
	2	10

Tests of Between-Subjects Effects

Dependent Variable: BACTERIA

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	695.048 ^a	2	347.524	152.276	.000
Intercept	34.808	1	34.808	15.252	.001
DAY	297.974	1	297.974	130.564	.000
LACTOSE	397.074	1	397.074	173.987	.000
Error	38.797	17	2.282		
Total	3279.494	20			
Corrected Total	733.846	19			

a. R Squared = .947 (Adjusted R Squared = .941)

Parameter Estimates

Dependent Variable: BACTERIA

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	7.550	.861	8.766	.000	5.733	9.367
DAY	2.729	.239	11.426	.000	2.225	3.233
[LACTOSE=1]	-8.912	.676	-13.190	.000	-10.337	-7.486
[LACTOSE=2]	0 ^a

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

and the second analysis produces the following ANOVA table instead...

(Contd.)

SPSS OUTPUT FOR BOX 6.3 WITH BOTH TYPES OF SUMS OF SQUARE (Contd.)

Tests of Between-Subjects Effects

Dependent Variable: BACTERIA

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	695.048 ^a	2	347.524	152.276	.000
Intercept	2545.648	1	2545.648	1115.434	.000
DAY	297.974	1	297.974	130.564	.000
LACTOSE	397.074	1	397.074	173.987	.000
Error	38.797	17	2.282		
Total	3279.494	20			
Corrected Total	733.846	19			

a. R Squared = .947 (Adjusted R Squared = .941)

Orthogonality is illustrated by the fact that the Type I and Type III SS are the same for both variables. Orthogonality may then be demonstrated more directly, by using DAY as the response variable.

SPSS COMMANDS FOR BOX 6.4

Illustrating orthogonality between continuous and categorical variables

Syntax `glm DAY by LACTOSE`
 `/print parameters`
 `/design LACTOSE.`

Menu route Analyze > General Linear Model > Univariate
 DAY → Dependent Variable
 LACTOSE → Fixed Factor(s)
 Model
 Custom
 Factors & Covariates → Build Terms → Model
 LACTOSE → LACTOSE
 Options
 Parameter estimates

SPSS OUTPUT FOR BOX 6.4

Illustrating orthogonality between continuous and categorical variables**General linear model****Between-Subjects Factors**

		N
LACTOSE	1	10
	2	10

Tests of Between-Subjects Effects

Dependent Variable: DAY

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.000 ^a	1	.000	.000	1.000
Intercept	180.000	1	180.000	81.000	.000
LACTOSE	.000	1	.000	.000	1.000
Error	40.000	18	2.222		
Total	220.000	20			
Corrected Total	40.000	19			

a. R Squared = .000 (Adjusted R Squared = -.056)

Parameter Estimates

Dependent Variable: DAY

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	3.000	.471	6.364	.000	2.010	3.990
[LACTOSE=1]	.000	.667	.000	1.000	-1.401	1.401
[LACTOSE=2]	0 ^a

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

6.4 Treating variables as continuous or categorical

Some variables may be legitimately treated as continuous or categorical. DAY in the *bacterial growth* dataset is an example of this.

SPSS COMMANDS FOR BOX 6.5

Analysing bacterial growth with DAY as a categorical variable

Syntax `glm BACTERIA by LACTOSE DAY`
 `/print parameters`
 `/design DAY LACTOSE.`

Menu route Analyze > General Linear Model > Univariate
 BACTERIA → Dependent Variable
 DAY LACTOSE → Fixed Factor(s)
 Model
 Custom
 Factors & Covariates → Build Terms → Model
 DAY → DAY
 LACTOSE → LACTOSE
 Options
 Parameter estimates

SPSS OUTPUT FOR BOX 6.5 **Analysing bacterial growth with DAY as a categorical variable**

General linear model

Between-Subjects Factors

		N
LACTOSE	1	10
	2	10
DAY	1	4
	2	4
	3	4
	4	4
	5	4

Tests of Between-Subjects Effects

Dependent Variable: BACTERIA

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	695.582 ^a	5	139.116	50.901	.000
Intercept	2545.648	1	2545.648	931.415	.000
DAY	298.508	4	74.627	27.305	.000
LACTOSE	397.074	1	397.074	145.284	.000
Error	38.263	14	2.733		
Total	3279.494	20			
Corrected Total	733.846	19			

a. R Squared = .948 (Adjusted R Squared = .929)

6.7 Exercises

Conservation and its influence on biomass

These data are stored in the *conservation* dataset.

SPSS COMMANDS FOR BOX 6.7 Conservation and biomass analysis. Execute once without and once with the bold commands	
Syntax	<pre>glm BIOMASS by CONS SOIL with ALT /print parameters /method sstype(1) /design CONS ALT SOIL.</pre>
Menu route	<p>Analyze > General Linear Model > Univariate</p> <p>BIOMASS → Dependent Variable</p> <p>ALT → Covariate(s)</p> <p>CONS SOIL → Fixed Factor(s)</p> <p><input type="text" value="Model"/></p> <p><input checked="" type="radio"/> Custom</p> <p>Factors & Covariates → Build Terms → Model</p> <p>CONS → CONS</p> <p>ALT → ALT</p> <p>SOIL → SOIL</p> <p>Sum of Squares: <input type="text" value="Type I ▼"/></p> <p><input type="text" value="Options"/></p> <p><input checked="" type="checkbox"/> Parameter estimates</p>

SPSS OUTPUT FOR BOX 6.7 **Conservation and biomass analysis****General linear model****Between-Subjects Factors**

		N
CONS	1	17
	2	33
SOIL	1	16
	2	18
	3	16

Tests of Between-Subjects Effects

Dependent Variable: BIOMASS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6.992 ^a	4	1.748	196.670	.000
Intercept	70.363	1	70.363	7916.338	.000
CONS	2.488E-02	1	2.488E-02	2.799	.101
ALT	4.427	1	4.427	498.103	.000
SOIL	.395	2	.198	22.238	.000
Error	.400	45	8.888E-03		
Total	166.311	50			
Corrected Total	7.392	49			

a. R Squared = .946 (Adjusted R Squared = .941)

(Contd.)

SPSS OUTPUT FOR BOX 6.7 (Contd.)

Parameter Estimates

Dependent Variable: BIOMASS

Parameter	B	Std. Error	t	Sig.	95% Confidence	
					Lower Bound	Upper Bound
Intercept	2.111	.034	62.385	.000	2.043	2.179
[CONS=1]	-4.89E-02	.029	-1.673	.101	-.108	9.963E-03
[CONS=2]	0 ^a
ALT	-2.91E-03	.000	-22.318	.000	-3.169E-03	-2.644E-03
[SOIL=1]	.231	.035	6.516	.000	.160	.302
[SOIL=2]	.145	.033	4.450	.000	7.925E-02	.210
[SOIL=3]	0 ^a

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

and the second analysis produces the following ANOVA table instead....

Tests of Between-Subjects Effects

Dependent Variable: BIOMASS

Source	Type I Sum of Squares	df	Mean	F	Sig.
Corrected Model	6.992 ^a	4	1.748	196.670	.000
Intercept	158.919	1	158.919	17879.574	.000
CONS	.718	1	.718	80.740	.000
ALT	5.879	1	5.879	661.464	.000
SOIL	.395	2	.198	22.238	.000
Error	.400	45	8.888E-03		
Total	166.311	50			
Corrected Total	7.392	49			

a. R Squared = .946 (Adjusted R Squared = .941)

Determinants of the grade point average

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