

6

Combining continuous and categorical variables

6.2 Combining continuous and categorical variables

To combine continuous and categorical variables, the word equation and variables are specified as before. In the menus, we distinguish the two kinds of variable by adding them to the 'Class' or 'Quantitative' pane. In the command line route, SAS assumes variables are continuous unless they are named in a Class statement.

Looking for a treatment for leprosy

Combining continuous and categorical variables is illustrated below with the *leprosy* dataset.

SAS COMMANDS FOR BOX 6.1 A treatment for leprosy

Commands `proc glm data=gandh.Chapter6;`
`class TREATMT;`
`model BACAFTER=BACBEF TREATMT / solution;`
`run;`

Menu route Statistics > Anova > Linear Models...

BACAFTER → Dependent

TREATMT → Class

BACBEF → Quantitative

Statistics

Parameter Estimates

Type I

62 Combining continuous and categorical variables

SAS OUTPUT FOR BOX 6.1 A treatment for leprosy					
The GLM Procedure					
Dependent Variable: BACAFTER					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	670.827478	223.609159	15.29	<.0001
Error	26	380.151814	14.621224		
Corrected Total	29	1050.979292			
	R-Square	Coeff Var	Root MSE	BACAFTER Mean	
	0.638288	38.72840	3.823771	9.873300	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
BACBEF	1	587.4800981	587.4800981	40.18	<.0001
TREATMT	2	83.3473802	41.6736901	2.85	0.0760
Source	DF	Type III SS	Mean Square	F Value	Pr > F
BACBEF	1	515.0147401	515.0147401	35.22	<.0001
TREATMT	2	83.3473802	41.6736901	2.85	0.0760
Parameter	Estimate	Standard Error	t Value	Pr > t	
Intercept	2.303341923 B	2.10276888	1.10	0.2834	
BACBEF	0.883070237	0.14879121	5.93	<.0001	
TREATMT 1	-3.905961455 B	1.73263338	-2.25	0.0328	
TREATMT 2	-3.041813293 B	1.71390586	-1.77	0.0876	
TREATMT 3	0.000000000 B	.	.	.	
NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.					

Once again the last treatment (level 3) has been aliased. 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:

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

This would give the following three equations:

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

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, but first we have a subtlety in SAS to reveal. In the command line route, we write the model formula directly and so we automatically specify the order of terms in the model. So far in the menu route, we have only said which variables are to be continuous ('Quantitative') and which are to be categorical ('Class'). How can we insist on a particular order of variables in the model, using the menu route? The box below shows that we must use the 'Model' button. This is our first use of two new conventions. 1) We name the pane or panes involved in a heading, and show indented below which actions are to be taken. The panes here are 'Effects in Model' and 'Independent'. The latter is a window showing the variables available, the former is the window in which we specify the order in which we wish the variables to be fitted. 2) If the actions involve moving a variable from one pane to the other, this is illustrated in the heading thus: 'Independent → Effects in Model'. We then state the button to be used (Add) to show how to move 'SEX' from one pane to another, thus 'SEX → Add'. From now on, we will often specify use of the Model button in the menu route.

The key point here is that we wish to fit WEIGHT as the first term in the model, and SEX as the second. To achieve this, this is the order in which they must appear in the 'Effects in model' pane which appears on using the 'Model' button. In this example, we illustrate what to do if the variables appear the other way around (with SEX first): that is, remove SEX and then add it again below WEIGHT.

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

```
Commands  proc glm data=gandh.Chapter6;
           class SEX;
           model FAT = WEIGHT SEX / solution;
           run;
```

```
Menu route  Statistics > Anova > Linear Models...

           FAT → Dependent
           SEX → Class
           WEIGHT → Quantitative
           Model
           Effects in model
           SEX → Remove
           Independent → Effects in model
           SEX → Add
           Statistics
            Parameter Estimates
            Type I
```

SAS OUTPUT FOR BOX 6.2 **Both weight and sex as explanatory variables for fat content**

The GLM Procedure

Dependent Variable: FAT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	177.4259728	88.7129864	34.62	<.0001
Error	16	40.9950799	2.5621925		
Corrected Total	18	218.4210526			

R-Square	Coeff Var	Root MSE	FAT Mean
0.812312	5.642489	1.600685	28.36842

Source	DF	Type I SS	Mean Square	F Value	Pr > F
WEIGHT	1	1.3282384	1.3282384	0.52	0.4819
SEX	1	176.0977344	176.0977344	68.73	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
WEIGHT	1	87.1049201	87.1049201	34.00	<.0001
SEX	1	176.0977344	176.0977344	68.73	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	9.058524204 B	3.00006269	3.02	0.0081
WEIGHT	0.217147050	0.03724247	5.83	<.0001
SEX 1	7.903750842 B	0.95337165	8.29	<.0001
SEX 2	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

6.3 Orthogonality in the context of continuous and categorical variables

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

SAS COMMANDS FOR BOX 6.3 **Bacterial growth at two levels of lactose**

Commands `proc glm data=gandh.Chapter6;`
`class LACTOSE;`
`model BACTERIA = DAY LACTOSE / solution;`
`run;`

Menu route Statistics > Anova > Linear Models...

BACTERIA → Dependent

LACTOSE → Class

DAY → Quantitative

Model

Effects in model

LACTOSE → Remove

Independent → Effects in model

LACTOSE → Add

Statistics

Parameter Estimates

Type I

SAS OUTPUT FOR BOX 6.3 Bacterial growth at two levels of lactose						
The GLM Procedure						
Dependent Variable: Bacteria						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	2	695.0482182	347.5241091	152.28	<.0001	
Error	17	38.7974808	2.2822048			
Corrected Total	19	733.8456989				
	R-Square	Coeff Var	Root MSE	Bacteria Mean		
	0.947131	13.39039	1.510697	11.28195		
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
Day	1	297.9740569	297.9740569	130.56	<.0001	
Lactose	1	397.0741613	397.0741613	173.99	<.0001	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
Day	1	297.9740569	297.9740569	130.56	<.0001	
Lactose	1	397.0741613	397.0741613	173.99	<.0001	
Parameter	Estimate	Standard Error	t Value	Pr > t		
Intercept	7.549650000 B	0.86122967	8.77	<.0001		
Day	2.729350000	0.23886213	11.43	<.0001		
Lactose 1	-8.911500000 B	0.67560414	-13.19	<.0001		
Lactose 2	0.000000000 B	.	.	.		
NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.						

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.

SAS COMMANDS FOR BOX 6.4 Illustrating orthogonality between continuous and categorical variables	
Commands	<pre>proc glm data=gandh.Chapter6; class LACTOSE; model DAY = LACTOSE / solution; run;</pre>
Menu route	Statistics > Anova > Linear Models... DAY → Dependent LACTOSE → Class <div style="border: 1px solid black; padding: 2px; display: inline-block;">Statistics</div> <input checked="" type="checkbox"/> Parameter Estimates <input checked="" type="checkbox"/> Type I

SAS OUTPUT FOR BOX 6.4 **Illustrating orthogonality between continuous and categorical variables**

The GLM Procedure

Dependent Variable: Day

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00000000	0.00000000	0.00	1.0000
Error	18	40.00000000	2.22222222		
Corrected Total	19	40.00000000			

R-Square	Coeff Var	Root MSE	Day Mean
0.000000	49.69040	1.490712	3.000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Lactose	1	0	0	0.00	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Lactose	1	0	0	0.00	1.0000

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	3.000000000 B	0.47140452	6.36	<.0001
Lactose 1	0.000000000 B	0.66666667	0.00	1.0000
Lactose 2	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

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.

SAS COMMANDS FOR BOX 6.5 **Analysing bacterial growth with DAY as a categorical variable**

Commands `proc glm data=gandh.Chapter6;`
`class DAY LACTOSE;`
`model BACTERIA = DAY LACTOSE / solution;`
`run;`

Menu route Statistics > Anova > Linear Models...
BACTERIA → Dependent
DAY LACTOSE → Class
DAY → Quantitative
Model
Effects in model
DAY → Remove
Independent → Effects in model
DAY → Add
Statistics
 Parameter Estimates
 Type I

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

The GLM Procedure

Dependent Variable: Bacteria

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	695.5823305	139.1164661	50.90	<.0001
Error	14	38.2633685	2.7330977		
Corrected Total	19	733.8456989			

R-Square	Coeff Var	Root MSE	Bacteria Mean
0.947859	14.65357	1.653208	11.28195

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Day	4	298.5081692	74.6270423	27.30	<.0001
Lactose	1	397.0741613	397.0741613	145.28	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Day	4	298.5081692	74.6270423	27.30	<.0001
Lactose	1	397.0741613	397.0741613	145.28	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	21.39700000 B	0.90549949	23.63	<.0001
Day 1	-10.93225000 B	1.16899481	-9.35	<.0001
Day 2	-8.46475000 B	1.16899481	-7.24	<.0001
Day 3	-5.86375000 B	1.16899481	-5.02	0.0002
Day 4	-3.03575000 B	1.16899481	-2.60	0.0211
Day 5	0.00000000 B	.	.	.
Lactose 1	-8.91150000 B	0.73933724	-12.05	<.0001
Lactose 2	0.00000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

6.7 Exercises

Conservation and its influence on biomass

These data are stored in the *conservation* dataset.

SAS COMMANDS FOR BOX 6.7 **Conservation and biomass analysis**

Commands `proc glm data=gandh.Chapter6;`
 `class CONS SOIL;`
 `model BIOMASS = CONS ALT SOIL / solution;`
 `run;`

Menu route Statistics > Anova > Linear Models...

BIOMASS → Dependent

ALT → Class

CONS SOIL → Quantitative

Effects in model

ALT SOIL → Remove

Independent → Effects in model

ALT → Add

SOIL → Add

Parameter Estimates

Type I

SAS OUTPUT FOR BOX 6.7 **Conservation and biomass analysis**

The GLM Procedure

Dependent Variable: BIOMASS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6.99223507	1.74805877	196.67	<.0001
Error	45	0.39997293	0.00888829		
Corrected Total	49	7.39220800			

R-Square	Coeff Var	Root MSE	BIOMASS Mean
0.945893	5.288182	0.094278	1.782800

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CONS	1	0.71764472	0.71764472	80.74	<.0001
ALT	1	5.87928103	5.87928103	661.46	<.0001
SOIL	2	0.39530932	0.19765466	22.24	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CONS	1	0.02487695	0.02487695	2.80	0.1013
ALT	1	4.42728568	4.42728568	498.10	<.0001
SOIL	2	0.39530932	0.19765466	22.24	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	2.110732461 B	0.03383375	62.39	<.0001
CONS 1	-0.048863422 B	0.02920750	-1.67	0.1013
CONS 2	0.000000000 B	.	.	.
ALT	-0.002906785	0.00013024	-22.32	<.0001
SOIL 1	0.230993900 B	0.03544796	6.52	<.0001
SOIL 2	0.144771860 B	0.03252949	4.45	<.0001
SOIL 3	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Determinants of the grade point average

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