

3

Models, parameters and GLMS

In SAS, the general linear model command (PROC GLM) that we used in Chapter 1 can be used to analyse models with both continuous and categorical explanatory variables, instead of using PROC REG for continuous variables. SAS will always assume that any explanatory variables used in the model formulae are continuous unless you specify otherwise.

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 'solution' option in the commands causes SAS to print the parameter estimates. Note that in the menu option HEIGHT is categorised as Quantitative rather than Class because it is continuous.

SAS COMMANDS FOR BOX 3.1 General linear models with continuous explanatory variables

Commands `proc glm data= gandh.Chapter3;`
`model VOLUME = HEIGHT / solution;`
`run;`

Menu route Statistics > Anova > Linear Models...

VOLUME → Dependent

HEIGHT → Quantitative

Statistics

Parameter estimates

This will produce the following output:

SAS OUTPUT FOR BOX 3.1 General Linear Model for the trees data set					
The GLM Procedure					
Dependent Variable: VOLUME					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2901.188867	2901.188867	16.16	0.0004
Error	29	5204.895004	179.479138		
Corrected Total	30	8106.083871			
	R-Square	Coeff Var	Root MSE	VOLUME Mean	
	0.357903	44.40356	13.39698	30.17097	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
HEIGHT	1	2901.188867	2901.188867	16.16	0.0004
Source	DF	Type III SS	Mean Square	F Value	Pr > F
HEIGHT	1	2901.188867	2901.188867	16.16	0.0004
Parameter	Estimate	Standard Error	t Value	Pr > t	
Intercept	-87.12361354	29.27312209	-2.98	0.0058	
HEIGHT	1.54334975	0.38386927	4.02	0.0004	

It can be seen that exactly the same information is given in the output as in the output from PROC REG in Chapter 2. 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 0.0004). The first ANOVA table is for the whole model. SAS also automatically provides both sequential and adjusted sums of squares for HEIGHT, which are called Type I and Type III, respectively. It does this in two separate ANOVA tables. (The meanings of these two forms of sum of squares is given in the main text). As there is only one explanatory variable in this model, the sum of squares for all three ANOVA tables is the same (2901.19). Finally, the coefficient table tells us *how* the variables are related ($VOLUME = -87.12 + 1.5433 \text{ HEIGHT}$).

Similarly, the commands for analysing the *fertiliser* data set are following:

SAS COMMANDS FOR BOX 3.2 **General linear models with categorical explanatory variables**

```

Commands  proc glm data= gandh.Chapter3;
           class FERTIL;
           model YIELD= FERTIL / solution;
           run;

```

Menu route Statistics > Anova > Linear Models...

YIELD → Dependent

FERTIL → Class

Statistics

Parameter estimates

This will give the following output:

SAS OUTPUT FOR BOX 3.2 **General linear model for the *fertiliser* data set**

The GLM Procedure

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	10.82274509	5.41137254	5.70	0.0086
Error	27	25.62214994	0.94896852		
Corrected Total	29	36.44489503			

R-Square	Coeff Var	Root MSE	YIELD Mean
0.296962	20.97804	0.974150	4.643667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
FERTIL	2	10.82274509	5.41137254	5.70	0.0086

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FERTIL	2	10.82274509	5.41137254	5.70	0.0086

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	4.487000060 B	0.30805333	14.57	<.0001
FERTIL 1	0.957999873 B	0.43565319	2.20	0.0366
FERTIL 2	-0.488000035 B	0.43565319	-1.12	0.2725
FERTIL 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.

Notice how the sum of squares for FERTIL appears in three places: (1) As Model sum of squares (because it is the only term in the model); (2) As FERTIL Type I SS and; (3) As FERTIL Type III SS. The adjusted and sequential sums of squares are equal again because FERTIL is the only and therefore the last term in the model.

The frightening warning at the bottom of the SAS 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. SAS'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 means for the three fertilisers, 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.

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

SAS COMMANDS FOR BOX 3.3 Obtaining least squares means for a categorical variable	
Commands	<pre>proc glm data=gandh.Chapter3; class FERTIL; model YIELD=FERTIL; lsmeans FERTIL; run;</pre>
Menu route	Statistics > Anova > Linear Models... YIELD → Dependent FERTIL → Class Means with LSMeans tab FERTIL → LS Mean

This will produce the following table in addition to the output shown in Box 3.2. The term 'least squares means' is used, as the data are fitted to the model by minimising the residual sum of squares.

SAS OUTPUT FOR BOX 3.3 Obtaining the group means for the fertiliser dataset

The GLM Procedure
Least Squares Means

FERTIL	YIELD LSMEAN
1	5.44499993
2	3.99900002
3	4.48700006

3.3 Turning the tables and creating datasets

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

The dataset Chapter3sim is available on the website. First import this dataset (reviving the `gandh` library if necessary). It has `YIELD` and `FERTIL` from the fertiliser dataset for comparison, and it also has additional columns whose values we will define. The parameters in the simulation are `K3`, `K1`, `K2` and `SIGMA`. These are represented here as full length variables, so we will assign values so that the value of `K3`, for example, is the same for every datapoint. Here are the commands (SAS commands for Box 3.3), available as `Chapter3sim.dat` on the website, or you may prefer to type them in.

SAS COMMANDS FOR BOX 3.3 Creating a dataset	
First group	<pre> data withdums; set gandh.Chapter3sim; dum1=(FERTIL=1); dum2=(FERTIL=2); run; </pre>
Second group	<pre> data withparameters; set withdums; k3=13.5; k1=1.2; k2=3.4; sigma=1.1; run; </pre>
Third group	<pre> data; set withparameters; noise=normal(0); y=k3+k1*dum1+k2*dum2+sigma*noise; run; proc glm; class FERTIL; model y=FERTIL / solution; run; </pre>

Chapter3sim.dat may be used as follows. Bring the Program Editor Window to the front (for example by View > Program Editor). Open this file in some suitable editor (such as Microsoft Word), and copy and paste the first group of commands into the Program Editor window, and select Run > Submit.

The first group of commands sets up the dummy variables, and you can look at them to see what their very simple structure is. SAS automatically saves the output in the work library, and in this you will find a new datasheet called 'withdums'. This will contain YIELD and FERTIL as the original chapter3sim file does, but also the two dummy variables. SAS's dummy variables are not the same as in the main text. 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, but is a consequence of SAS's method of aliasing.

Now copy and paste the second group of commands. Edit them to put the values you want to choose for the four parameters. Then choose Run > Submit. This sets the parameters. The new datasheet called 'withparameters' now to be found in the work library will contain values for the parameters K1, K2, K3 and SIGMA in addition to the information in 'withdums'.

Now copy-and-paste the third group of commands. This creates a temporary dataset with a set of random numbers in noise, calculates Y , and then performs a PROC GLM asking for the parameter estimates to be included in the output. The output window will come to the front and you will see the analysis of variance table and parameter estimates, scrolling as required. Record the information you want. The consequence of SAS's choice of aliasing can now be appreciated, namely that the constant, K_3 , is the mean we expect for $FERTIL=3$. The mean we expect for $FERTIL=1$ is K_3+K_1 and the mean we expect for $FERTIL=2$ is K_3+K_2 . As in the main text, if you make $SIGMA$ small enough, you will be able to confirm that this is indeed the case.

Now go back to the Program Editor window (for example by choosing View > Program Editor), and resubmit the commands by choosing Run > Submit. A new dataset with a fresh set of random variables will be created, and the output shown to you. Repeat as necessary until you have enough repetitions of the simulation. Each dataset is automatically saved in the work library, and called Data1, Data2 etc., for the duration of the session.

To change the parameters, all you need do is edit the values in the second group of commands to the parameter values you want, and resubmit the second and third groups of commands. An example set of output is shown below.

SAS OUTPUT FOR BOX 3.4 Recovering the parameter estimates					
The GLM Procedure					
Dependent Variable: Y					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	53.3948087	26.6974044	13.55	<.0001
Error	27	53.1884405	1.9699422		
Corrected Total	29	106.5832492			
	R-Square	Coeff Var	Root MSE	Y Mean	
	0.500968	9.433268	1.403546	14.87869	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
FERTIL	2	53.39480871	26.69740435	13.55	<.0001
Source	DF	Type III SS	Mean Square	F Value	Pr > F
FERTIL	2	53.39480871	26.69740435	13.55	<.0001
Parameter	Estimate	Standard Error	t Value	Pr > t	
Intercept	13.18777148 B	0.44384031	29.71	<.0001	
FERTIL 1	1.81156537 B	0.62768499	2.89	0.0076	
FERTIL 2	3.26117957 B	0.62768499	5.20	<.0001	
FERTIL 3	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.					

32 Models, parameters and GLMS

The means for the three levels of F can be calculated from the coefficients table. As mentioned earlier, SAS 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 SAS coefficients table as follows:

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

3.5 Exercises

How variability in the population will influence our analysis

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