

# 4

## Using more than one explanatory variable

### Commands and concepts introduced in this chapter

This chapter introduces models with more than one explanatory variable, and so leads to the concepts of statistical elimination, and adjusted, and sequential sums of squares. Both types of sum of squares are shown in the generic output given in the main text. In the main text, a single analysis of variance table shows the whole breakdown of sums of squares, into one error sum of square, and one sum of squares for each term in the model. With SAS, there are three ANOVA tables of concern to us. The first ANOVA table is always given, and shows the whole model versus the error. The second and third tables give the sequential and adjusted sums of squares, respectively, for each term in the model, and calls them Type I and Type III sums of squares respectively. The menu route produces automatically only the adjusted sums of squares, while the command line route produces both sequential and adjusted by default. Both types of sums of squares are automatically tested whenever they are shown. For the moment, we will focus on tests of the adjusted, Type III, sums of squares. Situations in which it would be more appropriate to use the sequential sum of squares, are explored and discussed in later chapters.

### 4.1 Why use more than one explanatory variable?

#### Leaping to the wrong conclusion

The first analysis in this example involves just one continuous explanatory variable, as in the previous chapter.

#### SAS COMMANDS FOR BOX 4.1 GLM with one explanatory variable

Commands `proc glm data=gandh.Chapter4;`  
`model AMA = HGT / SS3;`  
`run;`

Menu route Statistics > Anova > Linear Models...

AMA → Dependent

HGT → Quantitative

### 34 Using more than one explanatory variable

The option “/ SS3” instructs SAS to provide only the Adjusted Sums of Squares (Type III), which makes the command line and menu routes produce the same output. These commands give the following output:

SAS OUTPUT FOR BOX 4.1 Height explaining mathematical ability					
The GLM Procedure					
Dependent Variable: AMA					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	412.7743390	412.7743390	726.87	<.0001
Error	30	17.0364579	0.5678819		
Corrected Total	31	429.8107969			
	R-Square	Coeff Var	Root MSE	AMA Mean	
	0.960363	4.667481	0.753579	16.14531	
Source	DF	Type III SS	Mean Square	F Value	Pr > F
HGT	1	412.7743390	412.7743390	726.87	<.0001
Parameter	Estimate	Standard Error	t Value	Pr >  t	
Intercept	-42.82524588	2.19135073	-19.54	<.0001	
HGT	0.41180288	0.01527433	26.96	<.0001	

This analysis is then extended to include two explanatory variables. The additional variable is added to the word equation, separated by a space from the first explanatory variable. In this case both are continuous, which is the default in the command line route. Both must be added to the ‘Quantitative’ pane in the menu route. In order to see the Sequential (Type I) as well as the Adjusted (Type III) sums of squares, we must use the ‘Statistics...’ subdialog box in the menu route. This again makes the output from the two routes the same.

SAS COMMANDS FOR BOX 4.2 PROC GLM with two continuous explanatory variables	
Commands	<pre>proc glm data=gandh.Chapter4;   model AMA = YEARS HGT; run;</pre>
Menu route	Statistics > Anova > Linear Models... AMA → Dependent YEARS HGT → Quantitative Statistics <input checked="" type="checkbox"/> Type 1

This gives the following output:

SAS OUTPUT FOR BOX 4.2 <b>Years, not height explaining mathematical ability</b>					
The GLM Procedure					
Dependent Variable: AMA					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	422.5448613	211.2724307	850.80	<.0001
Error	29	7.2013763	0.2483233		
Corrected Total	31	429.7462377			
	R-Square	Coeff Var	Root MSE	AMA Mean	
	0.983243	3.086532	0.498321	16.14499	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
YEARS	1	422.5365257	422.5365257	1701.56	<.0001
HGT	1	0.0083357	0.0083357	0.03	0.8559
Source	DF	Type III SS	Mean Square	F Value	Pr > F
YEARS	1	9.86826762	9.86826762	39.74	<.0001
HGT	1	0.00833565	0.00833565	0.03	0.8559
Parameter	Estimate	Standard Error	t Value	Pr >  t	
Intercept	1.519547611	7.18118766	0.21	0.8339	
YEARS	-2.026026800	0.32139083	6.30	<.0001	
HGT	-0.012467632	0.06804918	-0.18	0.8559	

This illustrates SAS's useful convention in the command route of automatically providing tables with both the sequential and the adjusted sums of squares in the output. Both types of sum of square are tested, but in this case the adjusted or Type III sum of squares provides the most appropriate test. Later we shall come across examples where the sequential sum of squares proves to be more appropriate.

### Missing a significant relationship

This is a second example of the advantages of statistical elimination. The first analysis includes one explanatory variable.

36 Using more than one explanatory variable

SAS COMMANDS FOR BOX 4.3A <b>PROC GLM with one categorical explanatory variable</b>	
Commands	<pre>proc glm data=gandh.Chapter4;     class WATER;     model FINALHT = WATER; run;</pre>
Menu route	Statistics > Anova > Linear Models... FINALHT → Dependent WATER → Class

SAS OUTPUT FOR BOX 4.3A <b>Final height alone shows no differences between watering regimes</b>					
The GLM Procedure					
Dependent Variable: FINALHT					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12.89459000	4.29819667	1.97	0.1356
Error	36	78.46132000	2.17948111		
Corrected Total	39	91.35591000			
	R-Square	Coeff Var	Root MSE	FINALHT Mean	
	0.141147	26.54512	1.476307	5.561500	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
WATER	3	12.89459000	4.29819667	1.97	0.1356
Source	DF	Type III SS	Mean Square	F Value	Pr > F
WATER	3	12.89459000	4.29819667	1.97	0.1356

The second analysis includes a second explanatory variable.

SAS COMMANDS FOR BOX 4.3B **GLM with two explanatory variables**

```

Commands   proc glm data=gandh.Chapter4;
            class WATER;
            model FINALHT = WATER INITHT;
            run;

```

Menu route     Statistics > Anova > Linear Models...

FINALHT → Dependent

INITHT → Quantitative

WATER → Class

Statistics

Type 1

SAS OUTPUT FOR BOX 4.3B **Change in height is significantly different between watering regimes**

The GLM Procedure

Dependent Variable: FINALHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	91.16633661	22.79158415	4207.90	<.0001
Error	35	0.18957339	0.00541638		
Corrected Total	39	91.35591000			

	R-Square	Coeff Var	Root MSE	FINALHT Mean
	0.997925	1.323313	0.073596	5.561500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
WATER	3	12.89459000	4.29819667	793.55	<.0001
INITHT	1	78.27174661	78.27174661	14450.9	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
WATER	3	1.05193133	0.35064378	64.74	<.0001
INITHT	1	78.27174661	78.27174661	14450.9	<.0001

### 4.3 Two types of sum of squares

This section explores two possible consequences of statistical elimination.

#### Eliminating a third variable makes the second less informative

This is illustrated by comparing two models with one or two explanatory variables.

38 Using more than one explanatory variable

SAS COMMANDS FOR BOX 4.4A Length of right leg used to predict the weight of an individual	
Commands	<pre>proc glm data=gandh.Chapter4;       model WGHT = RLEG;       run;</pre>
Menu route	Statistics > Anova > Linear Models... WGHT → Dependent RLEG → Quantitative <div style="border: 1px solid black; display: inline-block; padding: 2px;">Statistics</div> <input checked="" type="checkbox"/> Type 1

SAS OUTPUT FOR BOX 4.4A Length of right leg predicts weight of an individual					
The GLM Procedure					
Dependent Variable: WGHT					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3627.670184	3627.670184	125.75	<.0001
Error	98	2827.099916	28.847958		
Corrected Total	99	6454.770100			
	R-Square	Coeff Var	Root MSE	WGHT Mean	
	0.562014	6.969661	5.371030	77.06300	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
RLEG	1	3627.670184	3627.670184	125.75	<.0001
Source	DF	Type III SS	Mean Square	F Value	Pr > F
RLEG	1	3627.670184	3627.670184	125.75	<.0001
Parameter	Estimate	Standard Error	t Value	Pr >  t	
Intercept	-3.731074382	7.22481246	-0.52	0.6067	
RLEG	1.008948511	0.08997309	11.21	<.0001	

SAS COMMANDS FOR BOX 4.4B **Using both RLEG and LLEG to predict weight**

```
Commands      proc glm data=gandh.Chapter4;
              model WGHT = RLEG LLEG;
              run;
```

Menu route Statistics > Anova > Linear Models...

WGHT → Dependent

RLEG LLEG → Quantitative

Statistics

Type 1

SAS OUTPUT FOR BOX 4.4B **Neither RLEG nor LLEG are significant predictors of weight**

The GLM Procedure

Dependent Variable: WGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3693.631807	1846.815903	64.88	<.0001
Error	97	2761.138293	28.465343		
Corrected Total	99	6454.770100			

	R-Square	Coeff Var	Root MSE	WGHT Mean
	0.572233	6.923286	5.335292	77.06300

Source	DF	Type I SS	Mean Square	F Value	Pr > F
RLEG	1	3627.670184	3627.670184	127.44	<.0001
LLEG	1	65.961623	65.961623	2.32	0.1312

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RLEG	1	83.33357120	83.33357120	2.93	0.0903
LLEG	1	65.96162276	65.96162276	2.32	0.1312

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	-1.991474451	7.26715540	-0.27	0.7846
RLEG	9.135643782	5.33933749	1.71	0.0903
LLEG	-8.147584906	5.35231225	-1.52	0.1312

The tests of the adjusted sums of squares lead to the rather counterintuitive conclusion that neither RLEG nor LLEG is a significant predictor of weight. The interpretation of these analyses is discussed in the main text.

**Eliminating a third variable makes the second more informative**

This example illustrates the converse situation.

**SAS COMMANDS FOR BOX 4.5A Using year of birth to predict a poet's age**

---

Commands      `proc glm data=gandh.Chapter4;`  
                   `model POETSAGE = BYEAR;`  
                   `run;`

---

Menu route      Statistics > Anova > Linear Models...  
                   POETSAGE → Dependent  
                   BYEAR → Quantitative  

Statistics

  
                    Type 1

**SAS OUTPUT FOR BOX 4.5A Age of poets cannot be predicted from birth date alone**

---

The GLM Procedure

Dependent Variable: POETSAGE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.161771	1.161771	0.00	0.9541
Error	10	3333.504896	333.350490		
Corrected Total	11	3334.666667			

R-Square	Coeff Var	Root MSE	POETSAGE Mean
0.000348	38.30326	18.25789	47.66667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BYEAR	1	1.16177051	1.16177051	0.00	0.9541

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BYEAR	1	1.16177051	1.16177051	0.00	0.9541

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	45.12881967	43.31074090	1.04	0.3220
BYEAR	0.00149777	0.02537088	0.06	0.9541

## SAS COMMANDS FOR BOX 4.5B Using years of birth and death to predict a poet's age

```
Commands      proc glm data=gandh.Chapter4;
               model POETSAGE = BYEAR DYEAR;
               run;
```

Menu route Statistics > Anova > Linear Models...

POETSAGE → Dependent

BYEAR DYEAR → Quantitative

Statistics

Type 1

## SAS OUTPUT FOR BOX 4.5B Age of poets can be accurately predicted from birth and death dates

The GLM Procedure

Dependent Variable: POETSAGE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3331.734308	1665.867154	5112.88	<.0001
Error	9	2.932358	0.325818		
Corrected Total	11	3334.666667			

	R-Square	Coeff Var	Root MSE	POETSAGE Mean
	0.999121	1.197492	0.570804	47.66667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BYEAR	1	1.161771	1.161771	3.57	0.0916
DYEAR	1	3330.572538	3330.572538	10222.2	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BYEAR	1	3299.663301	3299.663301	10127.3	<.0001
DYEAR	1	3330.572538	3330.572538	10222.2	<.0001

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	-0.311868922	1.42668447	-0.22	0.8318
BYEAR	-1.003901420	0.00997570	-100.63	<.0001
DYEAR	1.003685620	0.00992717	101.10	<.0001

In this case, the p-values based on the adjusted (Type III) sums of squares provide the most useful F ratio test.

No new principles arise in the urban fox example, or in the re-analysis of the *trees* dataset. We therefore move straight on to the exercises.

## 4.7 Exercises

### The cost of reproduction

The analysis involve models with one or two continuous explanatory variables.

SAS COMMANDS FOR BOX 4.11 Using reproductive effort to explain survival	
Commands	<pre>proc glm data=gandh.Chapter4;     model LLONGVTY = LEGGRATE; run;</pre>
Menu route	Statistics > Anova > Linear Models... LLONGVTY → Dependent LEGGRATE → Quantitative <div style="border: 1px solid black; padding: 2px; display: inline-block;">Statistics</div> <input checked="" type="checkbox"/> Type 1

SAS OUTPUT FOR BOX 4.11 GLM of survival against reproductive rate for <i>Drosophila sp</i>					
The GLM Procedure					
Dependent Variable: LLONGVTY					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7.73784119	7.73784119	5.83	0.0241
Error	23	30.50670425	1.32637845		
Corrected Total	24	38.24454544			
	R-Square	Coeff Var	Root MSE	LLONGVTY Mean	
	0.202325	67.06293	1.151685	1.717320	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
LEGGRATE	1	7.73784119	7.73784119	5.83	0.0241
Source	DF	Type III SS	Mean Square	F Value	Pr > F
LEGGRATE	1	7.73784119	7.73784119	5.83	0.0241
Parameter	Estimate	Standard Error	t Value	Pr >  t	
Intercept	1.769339461	0.23134172	7.65	<.0001	
LEGGRATE	0.281307922	0.11646778	2.42	0.0241	

## SAS COMMANDS FOR BOX 4.12 Using reproductive effort and size to explain survival

```
Commands      proc glm data=gandh.Chapter4;
               model LLONGVTY = LSIZE LEGGRATE;
               run;
```

Menu route    Statistics > Anova > Linear Models...

              LLONGVTY → Dependent

              LSIZE LEGGRATE → Quantitative

Type 1

SAS OUTPUT FOR BOX 4.12 GLM of survival against size and reproductive rate for *Drosophila sp*

The GLM Procedure

Dependent Variable: LLONGVTY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	29.57977095	14.78988547	37.55	<.0001
Error	22	8.66477449	0.39385339		
Corrected Total	24	38.24454544			

	R-Square	Coeff Var	Root MSE	LLONGVTY Mean
	0.773438	36.54400	0.627577	1.717320

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LSIZE	1	26.23971650	26.23971650	66.62	<.0001
LEGGRATE	1	3.34005445	3.34005445	8.48	0.0081

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LSIZE	1	21.84192975	21.84192975	55.46	<.0001
LEGGRATE	1	3.34005445	3.34005445	8.48	0.0081

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	1.681899213	0.12660861	13.28	<.0001
LSIZE	1.471877652	0.19764849	7.45	<.0001
LEGGRATE	-0.289926429	0.09955850	-2.91	0.0081

#### 44 Using more than one explanatory variable

To plot a graph similar to figure 4.10, you need to rank the data, according to size, and then create a new categorical explanatory variable called SIZEGRP. This can then be used to plot a graph in which the different size groups are represented by different symbols. These commands are conveniently done together in SAS, because the command line method takes the output from one command as the default input to the next. If you are in Analyst mode, then remember to move to Edit mode for the dataset, by choosing Edit > Mode > Edit. This is necessary to allow you to sort the data. As the data need sorting, this particular dataset has been saved in a separate file (*Drosophila*). This will prevent you sorting data for other datasets inadvertently!

SAS COMMANDS FOR FIGURE 4.10 **Sorting the data and plotting the graph**

```

Commands  proc sort data=gandh.Chapter4ex1;
           by LSIZE;

           data gandh.Chapter4ex1;
           set gandh.Chapter4ex1;
           label SIZEGRP = "Recoded Values of LSIZE";
           select(round(LSIZE, .000001));
             when (LSIZE <= -1.5) SIZEGRP = 1;
             when (-1.5 < LSIZE <= -0.5) SIZEGRP = 2;
             when (-0.5 < LSIZE <= 0.25) SIZEGRP = 3;
             when (0.25 < LSIZE <= 1) SIZEGRP = 4;
             when (1 < LSIZE <= 1.2) SIZEGRP = 5;
             when (1.2 < LSIZE) SIZEGRP = 6;
             otherwise;
           end;
           run;

           proc gplot;
           plot LLONGVTY*LEGRATE = SIZEGRP;
           run;

```

Menu route Data > Sort

LSIZE → Sort by

Data > Transform > Recode Ranges

Column to recode ▼

LSIZE

SIZEGRP → New column name

and new dialog box will appear

Enter 1 to 6 in the column 'New Value (numeric)' corresponding to the categories given in the commands above.

Graphs > Scatter Plot > Two-dimensional...

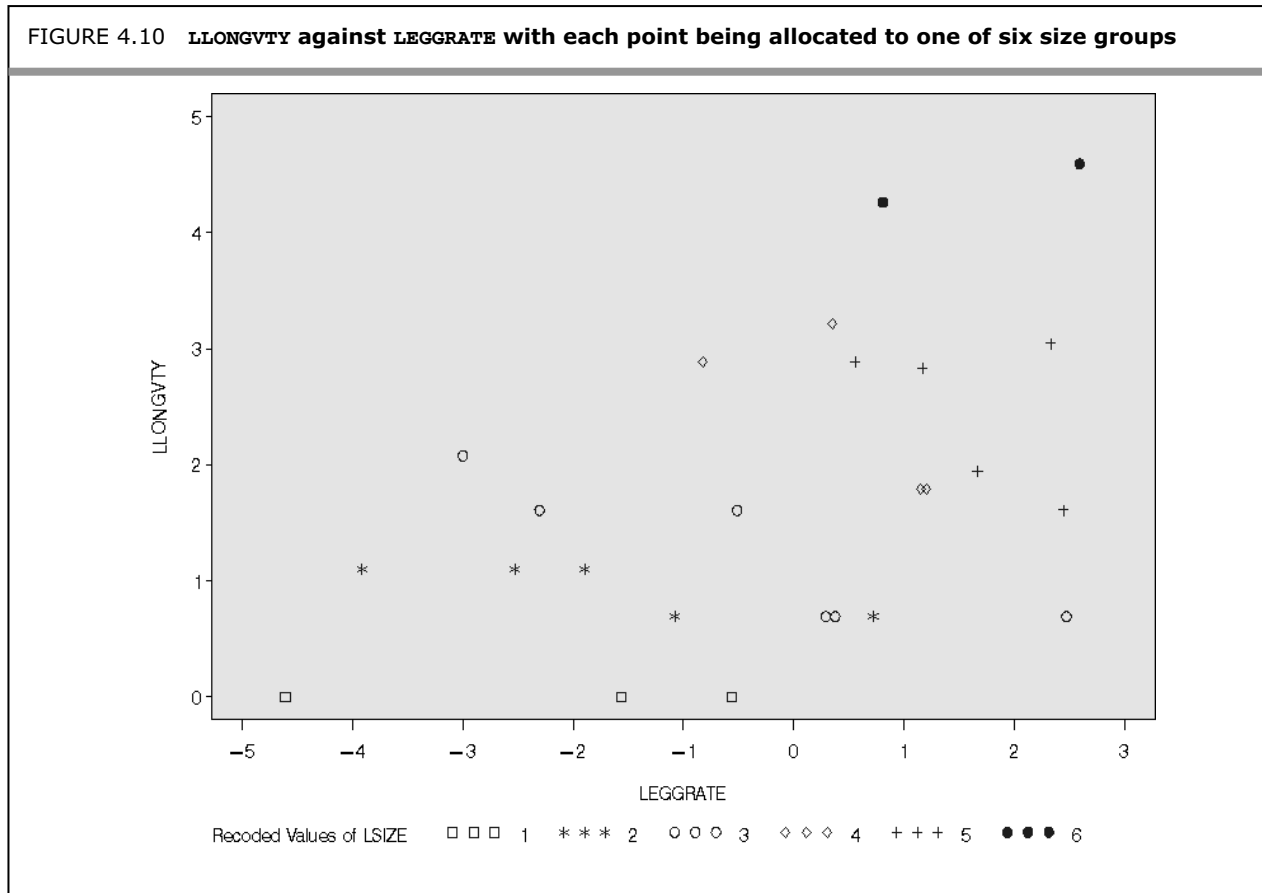
LEGRATE → X Axis

LLONGVTY → Y Axis

SIZEGRP → Class

46 Using more than one explanatory variable

This should produce the following graph:



**Investigating obesity**

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