

2

Regression

We will begin this chapter by talking about how you can import and save the files for a whole chapter, saving them in a library. On the companion website, the datasets are saved in two different formats. For SAS, we recommend that you use the format in which all datasets for one chapter are saved onto the same Excel worksheet (they are called SASdata2.xls, SASdata3.xls etc. with the number denoting the chapter number). The relevant file for this chapter is SASdata2.xls. You can then work through the examples for one chapter in one go, without having to import each data file separately.

First of all, in Analyst mode create a new library by choosing Tools > New library from the menus. In the new window, enter the library name (we suggest you use gandh for Grafen and Hails, to be consistent with this supplement). Leave the 'Engine' choice on default, and in the 'Path' pane, enter the directory where you would like to save your library. If you check the box 'Enable at setup', SAS will remember the library for subsequent sessions (otherwise you will need to repeat this procedure every time you open SAS). Once this is complete, go back to 'Initial' mode to import your data file as before (except this time choose SASdata2.xls from the directory in which you saved it after downloading from the web, and do not specify any particular worksheet within it). When you arrive at the 'Choose the SAS destination' window, choose your new library, gandh, from the drop down menu, and call the file chapter2. If you check your log window, you should see that gandh.chapter2 was successfully created.

If you use the command route, you refer to the data file as gandh.chapter2, and if you use the menu route in 'Analyst' mode, then you open the data file as before using File > Open by SAS name...

If you close the SAS session, and re-open at a later date, you will need to remind SAS of the library using Tools > New library, and in which directory it is stored, as before. However, once you have done this, you do not need to re-import the data. Resurrecting the library also revives the datasets within it. However, if you checked the box 'Enable at setup', as suggested above, you will not need to remind SAS of your new library.

2.4 Regression—an example

The trees data set

In this instance, we have used 'PROC REG' to investigate if HEIGHT is a significant predictor of VOLUME. If you are using menu routes, don't forget to first open the file using File > Open by SAS name...

SAS COMMANDS FOR BOX 2.1 Regression	
Commands	<pre>proc reg data=gandh.Chapter2; model VOLUME = HEIGHT; run;</pre>
Menu route	Statistics > Regression > Linear... VOLUME → Dependent HEIGHT → Explanatory

The output displays first the ANOVA table, and then the parameter estimates.

SAS OUTPUT FOR BOX 2.1 Analysis of the trees data set: regression					
The REG Procedure					
Model: MODEL1					
Dependent Variable: VOLUME					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2901.18887	2901.18887	16.16	0.0004
Error	29	5204.89500	179.47914		
Corrected Total	30	8106.08387			
	Root MSE	13.39698	R-Square	0.3579	
	Dependent Mean	30.17097	Adj R-Sq	0.3358	
	Coeff Var	44.40356			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-87.12361	29.27312	-2.98	0.0058
HEIGHT	1	1.54335	0.38387	4.02	0.0004

SAS produces two tables of output—the ANOVA table, and the coefficients table, just as in Box 2.1 of the main text. SAS also calculates R^2 and adjusted R^2 for you. (See chapter 11 for a discussion of adjusted R^2). It also provides you with ‘Root MSE’, the square root of the error mean square. In this case $\sqrt{179.5} = 13.4$. This is another measure of the ‘tightness’ of the data around the model. The ‘Dependent Mean’ and ‘Coeff Var’ are simple properties of the response variable, *VOLUME* (its mean and coefficient of variation).

2.6 Conclusions from a regression analysis

The next three analyses are also examples of simple linear regression. When doing successive analyses via the menu route, you can remove variables from a pane by highlighting them, and then clicking on the REMOVE tab (beneath the left hand pane which lists all possible variables for an analysis).

A strong relationship with little scatter

This analysis is with the *seeds* dataset.

SAS COMMANDS FOR BOX 2.2 Regression	
Commands	<pre>proc reg data=gandh.Chapter2; model SEEDWGHT = PLANDEN; run;</pre>
Menu route	Statistics > Regression > Linear... SEEDWGHT → Dependent PLANDEN → Explanatory

16 Regression

SAS OUTPUT FOR BOX 2.2 Analysis of the seeds data set: regression					
The REG Procedure					
Model: MODEL1					
Dependent Variable: SEEDWGHT					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10554	10554	111.45	<.0001
Error	18	1704.60743	94.70041		
Corrected Total	19	12259			
	Root MSE	9.73141	R-Square	0.8609	
	Dependent Mean	224.35000	Adj R-Sq	0.8532	
	Coeff Var	4.33760			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	311.89777	8.57377	36.38	<.0001
PLANDEN	1	-0.68773	0.06515	-10.56	<.0001

A weak relationship with lots of noise

This analysis is with the *scores* dataset.

SAS COMMANDS FOR BOX 2.3 Regression	
Commands	<pre>proc reg data=gandh.Chapter2; model MATHS = ESSAYS; run;</pre>
Menu route	Statistics > Regression > Linear... MATHS → Dependent ESSAYS → Explanatory

SAS OUTPUT FOR BOX 2.3 Analysis of the scores data set: regression						
The REG Procedure						
Model: MODEL1						
Dependent Variable: MATHS						
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	1	360.11905	360.11905	4.31	0.0715	
Error	8	668.28095	83.53512			
Corrected Total	9	1028.40000				
	Root MSE	9.13975	R-Square	0.3502		
	Dependent Mean	73.40000	Adj R-Sq	0.2689		
	Coeff Var	12.45198				
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	27.56667	22.26301	1.24	0.2507
ESSAYS	ESSAYS	1	0.65476	0.31535	2.08	0.0715

Small datasets and pet theories

This analysis is with the *rodent* dataset.

SAS COMMANDS FOR BOX 2.4 Regression	
Commands	<pre>proc reg data=gandh.Chapter2; model SPECIES2 = SPECIES1; run;</pre>
Menu route	<p>Statistics > Regression > Linear...</p> <p>SPECIES2 → Dependent</p> <p>SPECIES1 → Explanatory</p>

SAS OUTPUT FOR BOX 2.4 Analysis of the <i>rodent</i> data set: regression					
The REG Procedure					
Model: MODEL1					
Dependent Variable: SPECIES2					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	18.70130	18.70130	7.69	0.0694
Error	3	7.29870	2.43290		
Corrected Total	4	26.00000			
	Root MSE	1.55978	R-Square	0.7193	
	Dependent Mean	10.00000	Adj R-Sq	0.6257	
	Coeff Var	15.59776			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	14.51948	1.77308	8.19	0.0038
SPECIES1	1	-0.77922	0.28105	-2.77	0.0694

2.7 Unusual observations

Large residuals

In order to inspect residuals in the *trees* dataset, we need to repeat the analysis with some additional commands.

In the command line route, we create an extra dataset called 'score' in the 'work' library. (Datasets in the 'work' library are deleted when we quit SAS—those in other libraries are retained until we explicitly delete them.) All the existing variables in Chapter 2 are copied into the new dataset, and we request that the studentised residuals are also placed in it. 'Studentised residuals' is the term in SAS for what are called 'standardised residuals' in the main text. Note that we must put the additional command 'quit' after 'run', otherwise SAS will not create the new dataset 'score' until the next command is executed! In the menu route, much the same happens, except that the new dataset is created with a name you can't control (SCORE, and then on repeated analysis SCORE_i, where *i* is the next available integer) and it is stored in the library _PROJ_.

SAS COMMANDS TO SAVE THE RESIDUALS

```

Commands  proc reg data=gandh.Chapter2;
           model VOLUME = HEIGHT;
           output out=work.score student=_STUDENT;
           run;
           quit;

```

Menu route Statistics > Regression > Linear...

VOLUME → Dependent

HEIGHT → Explanatory

Save Data

Create and save diagnostic data

STUDENT → Add

Once the dataset has been created, we can view it as follows. The left hand pane has two tabs at the bottom. Click on the 'Explorer' tab to return to Explorer. This shows all the SAS datasets. You navigate up the file hierarchy by using the 'up one level' button in the toolbar, and down by double-clicking on libraries. If you have executed the commands in the editor window, you will need to navigate to the 'work' library, double-click on the 'score' dataset, and the dataset will appear in a spreadsheet-like format. You can view the residuals here. Of course, you may choose to save it to some other library, and call the file by some other name. If you used menus to save the residuals, you still find and view it in the same way, but now it will be in a new library called `_PROJ_`, and called SCORE. If you have done the analysis more than once, then the file name will be followed with the appropriate integer. In all cases, the new variable appears at the end (extreme right hand side), and will have the label 'Studentized residual' and the name '`_STUDENT`'.

The next step involves finding the extreme values and their positions.

SAS COMMANDS FOR BOX 2.5 Looking at large residuals (Point 31 of the trees dataset)

Commands `proc univariate data=work.score;`
`var _STUDENT;`
`run;`

Menu route *Step 1*

Double-click on the most recent 'Diagnostics Table' in the left hand pane of the Analyst window. You should be able to see the output dataset. It will look just like the original dataset, with the added column of residuals on the extreme right hand side.

Step 2

File > Save as by SAS Name...

and save it in the WORK library (click on WORK in the left hand pane, so selecting your library, and then type a file name in the Member Name box, and click on save). Remember the name you used!

Step 3

Close the diagnostics Table window (black cross in the top right hand corner). This should return you to the Analyst environment.

Step 4

File > Open by SAS Name...

and select the file you have just saved.

Step 5

Statistics > Descriptive > Distribution...

_STUDENT → Analysis

SAS OUTPUT FOR BOX 2.5 An excerpt from the PROC UNIVARIATE output on _STUDENT

```

The UNIVARIATE Procedure
Variable:  _STUDENT  (Studentized Residual)

      Extreme Observations

-----Lowest-----          -----Highest-----
      Value      Obs          Value      Obs
-1.648887         6          1.15786         29
-1.463937         5          1.25374         27
-1.443917        18          1.34311         26
-1.050040         9          1.67736         28
-0.817711         1          2.39116         31

```

SAS provides numerous statistics with the Univariate procedure—just an extract is reproduced here (you may have to scroll up or down the output window to find this part of the output). How these residuals are interpreted is explained in the text (the highest residual corresponds to the residual in Box 2.5 of the main text).

Influential points

The same routine is required to find influential points as standardised residuals, except that the key word is H instead of STUDENT, and the label is ‘Leverage’ instead of ‘Studentized Residuals’. H stands for ‘hat’, because the leverage values that indicate influence come from the diagonal elements of the so-called ‘hat matrix’.

SAS COMMANDS FOR INFLUENTIAL POINTS IN BOX 2.2

```

Commands  proc reg data=gandh.Chapter2;
           model SEEDWGHT = PLANDEN;
           output out=work.score H=_H;
           run;
           quit;

```

Menu route Statistics > Regression > Linear...

SEEDWGHT → Dependent

PLANDEN → Explanatory

Create and save diagnostic data

H → Add

View the contents of the score datasheets in the work or `_PROJ_` folders as before (in the work folder, the new Score spreadsheet will overwrite the old one, but in the `_PROJ_` folder it will now be called SCORE0 if you are doing this immediately after the preceding exercise), and use the univariate procedure on H to find the extreme values. This is done by repeating the SAS commands for Box 2.5 but replacing `_STUDENT` with `_H`. As a result, you should obtain the following output.

SAS OUTPUT FOR BOX 2.2		An excerpt from the PROC UNIVARIATE output			
The UNIVARIATE Procedure					
Variable: <code>_H</code> (Leverage)					
Extreme Observations					
-----Lowest-----		-----Highest-----			
Value	Obs	Value	Obs		
0.0504880	12	0.130186	10		
0.0509900	3	0.143595	11		
0.0526571	4	0.150263	16		
0.0530873	20	0.194074	14		
0.0542166	1	0.317780	8		

You can always use the explorer window (on the left) to navigate through the SAS environment to a results window. After the univariate procedure, there are a number of results windows and double clicking on each will produce a different set of measures—we have chosen the ‘extreme observations’ one here (or it may be called ‘moment and quartiles’ if you are using the menu route).

Datapoint number eight has the highest influence, with `_H=0.317780`. As discussed in the main text, one course of action is to remove the influential point and reanalyse the data.

The method of omitting datapoints varies with route. In the menu route, make sure you are in the Analyst environment, and follow `Data > Filter > Subset Data...` You then construct a rule for which datapoints to omit. The simplest way is to notice that datapoint eight is the only datapoint with `SEEDWGHT = 290`, and to proceed as follows. Click on `SEEDWGHT`, and a list of operators will appear. Click on `NE` (this stands for ‘not equals’). Then click on ‘Constant (enter value)’ in the list of columns to bring up a subsidiary dialog box. Type 290 into the ‘Enter Constant Value’ Box, and the click on ‘Apply’, then on ‘OK’. *Once you have run the analysis below, you should immediately go back and choose `Data > Filter > None`, otherwise you will be omitting the eighth datapoint from all further analyses on the Chapter2 dataset.*

With the command route, you need to include an extra statement in the PROC command that says ‘where `SEEDWGHT NE 290`’. The `NE` stands for ‘not equals’.

Thus the box shows how to perform the analysis with datapoint eight missing—but remember that in the menu route, the omission of datapoint eight has happened in a separate step.

SAS COMMANDS FOR BOX 2.6 Repeating analysis omitting the influential point

```

Commands  proc reg data=gandh.Chapter2;
           model SEEDWGHT = PLANDEN;
           where SEEDWGHT NE 290;
           run;

```

Menu route Statistics > Regression > Linear...

SEEDWGHT → Dependent

PLANDEN → Explanatory

SAS OUTPUT FOR BOX 2.6 Repeated analysis omitting the influential point

The REG Procedure
 Model: MODEL1
 Dependent Variable: SEEDWGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6245.79702	6245.79702	71.94	<.0001
Error	17	1475.99245	86.82309		
Corrected Total	18	7721.78947			

Root MSE	9.31789	R-Square	0.8089
Dependent Mean	220.89474	Adj R-Sq	0.7976
Coeff Var	4.21825		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	302.90982	9.90326	30.59	<.0001
PLANDEN	1	-0.62431	0.07361	-8.48	<.0001

2.10 Exercises**Does weight mean fat?**

The commands required for this exercise are the regression commands introduced in Chapter 2.

SAS COMMANDS FOR BOX 2.7 Analysis of reduced fat dataset	
Commands	<pre>proc reg data=gandh.Chapter2; model FAT = WEIGHT; run;</pre>
Menu route	Statistics > Regression > Linear... FAT → Dependent WEIGHT → Explanatory

These produce the following output:

SAS OUTPUT FOR BOX 2.7 Analysis of reduced fat dataset					
The REG Procedure Model: MODEL1 Dependent Variable: FAT					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.32824	1.32824	0.10	0.7510
Error	17	217.09281	12.77017		
Corrected Total	18	218.42105			
	Root MSE	3.57354	R-Square	0.0061	
	Dependent Mean	28.36842	Adj R-Sq	-0.0524	
	Coeff Var	12.59688			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	26.88558	4.67036	5.76	<.0001
WEIGHT	1	0.02069	0.06414	0.32	0.7510

Dioecious trees

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