

2

Regression

2.4 Regression—an example

The trees dataset

In this instance, we have used the ‘Regress’ command, to investigate if HEIGHT is a significant predictor of VOLUME. ‘Regress’ can do many, more complex operations, some of which we will encounter in later chapters. This example also illustrates the need to use the cumbersome subcommand ‘/method=enter’ just to specify the x-variable in the syntax route.

The output displays first a model summary, then the ANOVA table, and finally the coefficients table displaying the information needed for the fitted value equation.

SPSS COMMANDS FOR BOX 2.1 Regression	
Syntax	regression /dependent VOLUME /method=enter HEIGHT.
Menu route	Analyze > Regression > Linear VOLUME → Dependent HEIGHT → Independent(s)

SPSS OUTPUT FOR BOX 2.1

Analysis of variance and fitted value tables for the *trees* data set**Regression****Variables Entered/Removed^b**

Model	Variable Entered	Variables Removed	Method
1	HEIGHT ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: VOLUME

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.598 ^a	.358	.336	13.3970

a. Predictors: (Constant), HEIGHT

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2901.189	1	2901.189	16.164	.000 ^a
	Residual	5204.895	29	179.479		
	Total	8106.084	30			

a. Predictors: (Constant), HEIGHT

b. Dependent Variable: VOLUME

Coefficient^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-87.124	29.273		-2.976	.006
	HEIGHT	1.543	.384	.598	4.021	.000

a. Dependent Variable: VOLUME

In the model summary, SPSS calculates R , R^2 and adjusted R^2 for you. (See chapter 11 for a discussion of adjusted R^2). It also provides you with the “Standard Error of the Estimate”, SPSS’s peculiar name for the square root of the error mean square. In this case $\sqrt{179.479} = 13.397$. This is another measure of the ‘tightness’ of the data around the model. These points are discussed in more detail later in chapter 2.

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 arrow to the left of the Dependent or Independent(s) pane. This will move them back to the source pane.

A strong relationship with little scatter

This analysis is with the *seeds* dataset.

SPSS COMMANDS FOR BOX 2.2 Regression	
Syntax	<pre> regression /dependent SEEDWGHT /method=enter PLANDEN . </pre>
Menu route	<pre> Analyze > Regression > Linear SEEDWGHT → Dependent PLANDEN → Independent(s) </pre>

SPSS OUTPUT FOR BOX 2.2 **Regression analysis of SEEDWGHT against PLANDEN****Regression****Variables Entered/Removed^b**

Model	Variable Entered	Variables Removed	Method
1	PLANDEN ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: SEEDWGHT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.928 ^a	.861	.853	9.731

a. Predictors: (Constant), PLANDEN

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regressio	10553.943	1	10553.943	111.446	.000 ^a
	Residua	1704.607	18	94.700		
	Total	12258.550	19			

a. Predictors: (Constant), PLANDEN

b. Dependent Variable: SEEDWGHT

Coefficient^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.	Beta		
1	(Constant)	311.898	8.574		36.378	.000
	PLANDEN	-.688	.065	-.928	-10.557	.000

a. Dependent Variable: SEEDWGHT

A weak relationship with lots of noise

This analysis is with the *scores* dataset.

SPSS COMMANDS FOR BOX 2.3 Regression	
Syntax	<pre>regression /dependent MATHS /method=enter ESSAYS.</pre>
Menu route	Analyze > Regression > Linear MATHS → Dependent ESSAYS → Independent(s)

SPSS OUTPUT FOR BOX 2.3 **Regression analysis of MATHS and ESSAYS****Regression****Variables Entered/Removed^b**

Model	Variables Entered	Variables Removed	Method
1	ESSAYS ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: MATHS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.592 ^a	.350	.269	9.140

a. Predictors: (Constant), ESSAYS

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	360.119	1	360.119	4.311	.072 ^a
	Residual	668.281	8	83.535		
	Total	1028.400	9			

a. Predictors: (Constant), ESSAYS

b. Dependent Variable: MATHS

Coefficient^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	27.567	22.263		1.238	.251
	ESSAYS	.655	.315	.592	2.076	.072

a. Dependent Variable: MATHS

Small datasets and pet theories

This analysis is with the *rodent* dataset.

SPSS COMMANDS FOR BOX 2.4 Regression	
Syntax	<pre>regression /dependent SPECIES2 /method=enter SPECIES1.</pre>
Menu route	Analyze > Regression > Linear SPECIES2 → Dependent SPECIES1 → Independent(s)

SPSS OUTPUT FOR BOX 2.4 **Regression analysis of SPECIES1 and SPECIES2****Regression****Variables Entered/Removed^b**

Model	Variables Entered	Variables Removed	Method
1	SPECIES1 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: SPECIES2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.848 ^a	.719	.626	1.560

a. Predictors: (Constant), SPECIES1

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18.701	1	18.701	7.687	.069 ^a
	Residual	7.299	3	2.433		
	Total	26.000	4			

a. Predictors: (Constant), SPECIES1

b. Dependent Variable: SPECIES2

Coefficient^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	14.519	1.773		8.189	.004
	SPECIES1	-.779	.281	-.848	-2.773	.069

a. Dependent Variable: SPECIES2

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. The two methods (syntax and menus) are not exactly equivalent, as we shall see.

SPSS COMMANDS FOR BOX 2.5 Point 31 of the <i>trees</i> dataset	
Syntax	<pre> regression /dependent VOLUME /method=enter HEIGHT /casewise plot(sresid) outliers(2). </pre>
Menu route	Analyze > Regression > Linear VOLUME → Dependent HEIGHT → Independent(s) <div style="border: 1px solid black; padding: 2px; display: inline-block; margin-left: 20px;">Statistics</div> <input checked="" type="checkbox"/> Casewise diagnostics <input type="radio"/> Outliers outside 2 → standard deviations

In addition to the usual output we saw earlier, the syntax route produces two extra tables:

SPSS OUTPUT FOR BOX 2.5 **Via syntax route: Point 31 of the trees dataset****Casewise Diagnostics^a**

Case Number	Stud. Residual	VOLUME	Predicted Value	Residual
31	2.391	77.0	47.148	29.852

a. Dependent Variable: VOLUME

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	10.107	47.148	30.171	9.8339	31
Std. Predicted Value	-2.040	1.726	.000	1.000	31
Standard Error of Predicted Value	2.4062	5.5401	3.2781	.9279	31
Adjusted Predicted Value	9.304	47.955	30.090	9.9441	31
Residual	-21.274	29.852	.000	13.1718	31
Std. Residual	-1.588	2.228	.000	.983	31
Stud. Residual	-1.649	2.391	.003	1.022	31
Deleted Residual	-22.937	34.376	.081	14.2329	31
Stud. Deleted Residual	-1.702	2.622	.012	1.049	31
Mahal. Distance	.000	4.163	.968	1.177	31
Cook's Distance	.000	.433	.041	.080	31
Centered Leverage Value	.000	.139	.032	.039	31

a. Dependent Variable: VOLUME

The casewise diagnostics table notifies us of the same datapoint as in the main text, and with the same residual of 2.39. Notice the keyword “sresids” in the command line — this means “studentized residuals”, and is SPSS’s name for what we call “standardized residuals” in the main text. We use the command outliers (2) to request that SPSS picks out those points for which their standardised residuals are greater than 2 (or less than -2).

The menu route produces the following two tables instead:

SPSS OUTPUT FOR BOX 2.5 Via menu route: Point 31 of the <i>trees</i> dataset					
Casewise Diagnostics^a					
Case Number	Std. Residual	VOLUME	Predicted Value	Residual	
31	2.228	77.0	47.148	29.852	
a. Dependent Variable: VOLUME					
Residuals Statistics^a					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	10.107	47.148	30.171	9.8339	31
Residual	-21.274	29.852	.000	13.1718	31
Std. Predicted Value	-2.040	1.726	.000	1.000	31
Std. Residual	-1.588	2.228	.000	.983	31
a. Dependent Variable: VOLUME					

Here the same datapoint is picked out, but the residual is 2.28. This is what SPSS calls a standardized residual, and is simply calculated by dividing the residual itself by the square root of the Error Mean Square (so, in this case $2.28 = 29.852 / \sqrt{179.479}$). The second table produced gives summary information about a number of 'temporary variables' held during the SPSS calculation. We will not refer to this table again, and show it here only because it does actually appear in the output (and can't be prevented). Indeed, we will omit it to save space in later output from REGRESSION.

The residuals produced by the syntax route are consistent with the main text. The confusion over the naming of residuals between different statistical packages is considerable, and we show below a table comparing the terminology used in the main text with SPSS.

Comparing the names of residuals in SPSS and the main text	
Main text	SPSS
Standardised residual	Studentized residual
<< no equivalent >>	Studentized deleted
<< no equivalent >>	Standardized residual, i.e. just
	$\frac{\text{residual}}{\sqrt{\text{Error Mean Square}}}$

Influential points

Observations can be unusual because their residual is very high or low, as discussed in the previous section. But they can also be unusual because their x-values are very far from the ‘centre of gravity’ of the x-values of other datapoints. These datapoints are called ‘influential’ because they have a much greater role in determining the regression coefficients than more ‘typical’ datapoints. SPSS does not automatically draw our attention to influential datapoints, but we can easily ask for a measure of influence (called leverage) to be saved into a variable for us.

The commands are as follows: With the syntax route, we can choose the name of the variable (and we choose `PLANTLEV`). With menus, the name is chosen automatically and will be `LEV_1` if the name is not already being used. (More generally it is `LEV_k`, where k is the smallest positive integer so that `LEV_k` is not a name already being used.)

SPSS commands to identify influential points	
Syntax	<pre> regression /dependent SEEDWGHT /method=enter PLANDEN /save lever(PLANTLEV) . </pre>
Menu route	<p>Analyze > Regression > Linear</p> <p>SEEDWGHT → Dependent</p> <p>PLANDEN → Independent(s)</p> <div style="border: 1px solid black; display: inline-block; padding: 2px 5px;">Save</div> <p><input checked="" type="checkbox"/> Leverage values</p>

These commands would give you the same output as in SPSS output for Box 2.2. However, the leverage values can be inspected in the datasheet where you will now find them on the extreme righthand side. One view is that a leverage is high enough to be worth investigating if it is higher than $0.99-1/n$, or if it is higher than $(3p-1)/n$, where p is the number of parameters (including the constant), and n is the number of datapoints. In this case, $p=2$ (slope and intercept) and $n=20$, so leverages greater than 0.25 would be considered unusual. Only one datapoint has a leverage value greater than 0.25, and that is datapoint 8, whose value is 0.26778. Thus we would learn that datapoint 8 has an x-value far from the ‘centre of gravity’ of datapoints, and has a strong influence on the estimates of the coefficients.

As discussed in the main text, one course of action is to remove the influential point and reanalyse the data, as shown below. Removing datapoints is an important technique in practice, and so we now show how to do it in this simple case. The syntax route is long and cumbersome, and will not be used for this step. Make sure the Data window is at the front. Then choose Data > Select Cases..., and select the radio button 'If condition is satisfied', then click on the 'If...' button. This takes you to a calculator-like window. Select the leverage column you created in a previous step - it will have its label as 'Centered Leverage Value' and its name will be like 'lev_1'. Then click on the '<' button on the calculator keypad, followed by 0.25. This tells SPSS to include cases if their leverage value is less than 0.25. The unusual datapoint's value is greater than this, and will be the only one omitted. Now click on Continue. *This selection of datapoints continues for all analyses with this dataset until you actively change your choice, not just analyses with these variables. It also affects plots. It is therefore important to undo the selection as soon as possible. You do this by going back to the same menu route, and choosing 'Reset' in the main dialog box.*

The analysis of Box 2.2 can then be repeated, producing the output below:

SPSS OUTPUT FOR BOX 2.7 Repeated analysis omitting the influential point						
Regression						
Variables Entered/Removed^b						
Model	Variables Entered	Variables Removed	Method			
1	PLANDEN ^a	.	Enter			
a. All requested variables entered.						
b. Dependent Variable: SEEDWGHT						
Model Summary^b						
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate		
1	.899 ^a	.809	.798	9.318		
a. Predictors: (Constant), PLANDEN						
b. Dependent Variable: SEEDWGHT						
ANOVA^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6245.797	1	6245.797	71.937	.000 ^a
	Residual	1475.992	17	86.823		
	Total	7721.789	18			
a. Predictors: (Constant), PLANDEN						
b. Dependent Variable: SEEDWGHT						
Coefficients^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	302.910	9.903		30.587	.000
	PLANDEN	-.624	.074	-.899	-8.482	.000
a. Dependent Variable: SEEDWGHT						

2.10 Exercises

Does weight mean fat?

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

SPSS COMMANDS FOR BOX 2.7 Analysis of reduced fat dataset	
Syntax	<pre>regression /dependent FAT /method=enter WEIGHT.</pre>
Menu route	<pre>Analyze > Regression > Linear FAT → Dependent WEIGHT → Independent(s)</pre>

These produce the following output:

SPSS OUTPUT FOR BOX 2.7 Analysis of reduced fat dataset						
Regression						
Variables Entered/Removed^b						
Model	Variables Entered	Variables Removed	Method			
1	WEIGHT ^a	.	Enter			
a. All requested variables entered.						
b. Dependent Variable: FAT						
Model Summary						
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate		
1	.078 ^a	.006	-.052	3.574		
a. Predictors: (Constant), WEIGHT						
ANOVA^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.328	1	1.328	.104	.751 ^a
	Residual	217.093	17	12.770		
	Total	218.421	18			
a. Predictors: (Constant), WEIGHT						
b. Dependent Variable: FAT						
Coefficient^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.	Beta		
1	(Constant)	26.886	4.670		5.757	.000
	WEIGHT	2.069E-02	.064	.078	.323	.751
a. Dependent Variable: FAT						

Dioecious trees

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