

# 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.

#### MINITAB COMMANDS FOR BOX 2.1 Regression

Commands	Regress VOLUME 1 HEIGHT
Menu route	Stat > Regression > Regression VOLUME → Response HEIGHT → Predictors

The output displays first the fitted value equation, followed by the coefficients table, and below this the ANOVA table.

#### MINITAB OUTPUT FOR BOX 2.1 Analysis of the *trees* dataset: regression

##### Regression Analysis: VOLUME versus HEIGHT

The regression equation is  
VOLUME = - 87.1 + 1.54 HEIGHT

Predictor	Coef	SE Coef	T	P
Constant	-87.12	29.27	-2.98	0.006
HEIGHT	1.5433	0.3839	4.02	0.000

S = 13.40      R-Sq = 35.8%      R-Sq (adj) = 33.6%

##### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2901.2	2901.2	16.16	0.000
Residual Error	29	5204.9	179.5		
Total	30	8106.1			

## Unusual Observations

Obs	HEIGHT	VOLUME	Fit	SE Fit	Residual	St Resid
31	87.0	77.00	47.15	4.86	29.85	2.39R

R denotes an observation with a large standardized residual

Minitab 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  $s$ , 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. Finally, Minitab flags up any unusual observations, indicating in this instance that point 31 is an outlier (in other words, it is more than 2 standardised residuals away from its expected value). 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.

### A strong relationship with little scatter

This analysis is with the *seeds* dataset.

#### MINITAB COMMANDS FOR BOX 2.2 Regression

Commands	Regress SEEDWGHT 1 PLANDEN
Menu route	Stat > Regression > Regression SEEDWGHT → Response PLANDEN → Predictors

#### MINITAB OUTPUT FOR BOX 2.2 Analysis of the seeds dataset: regression

##### Regression Analysis: SEEDWGHT versus PLANDEN

The regression equation is

$$\text{SEEDWGHT} = 312 - 0.688 \text{ PLANDEN}$$

Predictor	Coef	SE Coef	T	P
Constant	311.898	8.574	36.38	0.000
PLANDEN	-0.68773	0.06515	-10.56	0.000

S = 9.731

R-Sq = 86.1%

R - Sq(adj) = 85.3%

Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	1	10554	10554	111.45	0.000	
Residual Error	18	1705	95			
Total	19	12259				

  

Unusual Observations						
Obs	PLANDEN	SEEDWGHT	Fit	SE Fit	Residual	St Resid
8	50	290.00	277.51	5.49	12.49	1.55 X

X denotes an observation whose X value gives it large influence.

### A weak relationship with lots of noise

This analysis is with the *scores* dataset.

MINITAB COMMANDS FOR BOX 2.3 Regression	
Commands	Regress MATHS 1 ESSAYS
Menu route	Stat > Regression > Regression MATHS → Response ESSAYS → Predictors

MINITAB OUTPUT FOR BOX 2.3 Analysis of the scores dataset: regression					
<b>Regression Analysis: MATHS versus ESSAYS</b>					
The regression equation is					
MATHS = 27.6 + 0.655 ESSAYS					
Predictor	Coef	SE Coef	T	P	
Constant	27.57	22.26	1.24	0.251	
ESSAYS	0.6548	0.3154	2.08	0.072	
S = 9.140	R-Sq = 35.0%	R-Sq (adj) = 26.9%			
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	360.12	360.12	4.31	0.072
Residual Error	8	668.28	83.54		
Total	9	1028.40			

### Small datasets and pet theories

This analysis is with the *rodents* dataset.

**MINITAB COMMANDS FOR BOX 2.4 Regression**

Commands	Regress SPECIES2 1 SPECIES1
Menu route	Stat > Regression > Regression SPECIES2 → Response SPECIES1 → Predictors

**MINITAB OUTPUT FOR BOX 2.4 Analysis of the rodents dataset: regression****Regression Analysis: SPECIES2 versus SPECIES1**

The regression equation is  
 $SPECIES2 = 14.5 - 0.779 SPECIES1$

Predictor	Coef	SE Coef	T	P
Constant	14.519	1.773	8.19	0.004
SPECIES1	-0.7792	0.2811	-2.77	0.069

S = 1.560      R-Sq = 71.9%      R-Sq(adj) = 62.6%

**Analysis of Variance**

Source	DF	SS	MS	F	P
Regression	1	18.701	18.701	7.69	0.069
Residual Error	3	7.299	2.433		
Total	4	26.00			

**2.7 Unusual observations****Large residuals**

Two of the datasets examined so far have produced additional output after the Analysis of Variance table. In these cases, Minitab is drawing our attention to certain features of the dataset. For example, in this case, the data point is more than two standardised residuals away from the expected value.

**MINITAB OUTPUT FOR BOX 2.5 Minitab's warning of a large residual (taken from Box 2.1)****Unusual Observations**

Obs	HEIGHT	VOLUME	Fit	SE Fit	Residual	St Resid
31	87.0	77.00	47.15	4.86	29.85	2.39R

R denotes an observation with a large standardized residual

## Influential points

In the *seeds* dataset (see Box 2.2), another type of unusual observation occurs. Datapoint number 8 is picked out, even though its standardised residual is only 1.55, because it is influential.

### MINITAB OUTPUT FOR BOX 2.6 Minitab's warning of an influential point (taken from Box 2.2)

#### Unusual Observations

Obs	PLANDEN	SEEDWGHT	Fit	SE Fit	Residual	St Resid
8	50	290.00	277.51	5.49	12.49	1.55 X

X denotes an observation whose X value gives it large influence.

Again, the interpretation of this warning is given in the main text. Minitab provides these warnings of large residuals and influential points automatically—no extra commands are required. As discussed in the main text, one course of action is to remove the influential point and reanalyse the data, as shown below. To remove a datapoint, select the cells, and choose Edit > Delete Cells. Remember not to save the dataset, or you will have to obtain a fresh copy to get the deleted datapoint back.

### MINITAB OUTPUT FOR BOX 2.7 Repeated analysis omitting the influential point

#### Regression Analysis: SEEDWGHT versus PLANDEN

The regression equation is

$$\text{SEEDWGHT} = 303 - 0.624 \text{ PLANDEN}$$

Predictor	Coef	SE Coef	T	P
Constant	302.910	9.903	30.59	0.000
PLANDEN	-0.62431	0.07361	-8.48	0.000

S = 9.318      R-Sq = 80.9%      R-Sq (adj) = 79.8%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	6245.8	6245.8	71.94	0.000
Residual Error	17	1476.0	86.8		
Total	18	7721.8			

#### Unusual Observations

Obs	PLANDEN	SEEDWGHT	Fit	SE Fit	Residual	St Resid
5	115	250.00	231.11	2.45	18.89	2.10R

R denotes an observation with a large standardized residual

## 2.10 Exercises

### Does weight mean fat?

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

MINITAB COMMANDS FOR BOX 2.7 Analysis of reduced fat dataset	
Commands	Regress FAT 1 WEIGHT
Menu route	Stat > Regression > Regression FAT → Response WEIGHT → Predictors

These produce the following output:

MINITAB OUTPUT FOR BOX 2.7 Analysis of reduced fat dataset					
<b>Regression Analysis: FAT versus WEIGHT</b>					
The regression equation is					
FAT = 26.9 + 0.0207 WEIGHT					
Predictor	Coef	SE Coef	T	P	
Constant	26.886	4.670	5.76	0.000	
WEIGHT	0.02069	0.06414	0.32	0.751	
S = 3.574	R-Sq = 0.6%	R-Sq (adj) = 0.0%			
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	1.33	1.33	0.10	0.751
Residual Error	17	217.09	12.77		
Total	18	218.42			

### Dioecious trees

See Minitab output for this exercise in the answers to exercises in Chapter 14.