1. Exercises

Melons

An experiment was performed to compare four melon varieties. It was designed so that each variety was grown in six plots–but two plots growing variety 3 were accidentally destroyed. The data are plotted in Fig. 1.15, and can be found in the *melons* dataset under the variables YIELDM and VARIETY.

Table 1.7 shows some summary statistics and an ANOVA table produced from these data.

(1) What is the null hypothesis in this case?

(2) What conclusions would you draw from the analysis in Table 1.7?

(3) How would you summarise the information provided by the data about the amount of error variation in the experiment?

(4) Calculate the standard error of the mean for all four varieties.

(5) How would you summarise and present the information from this analysis?

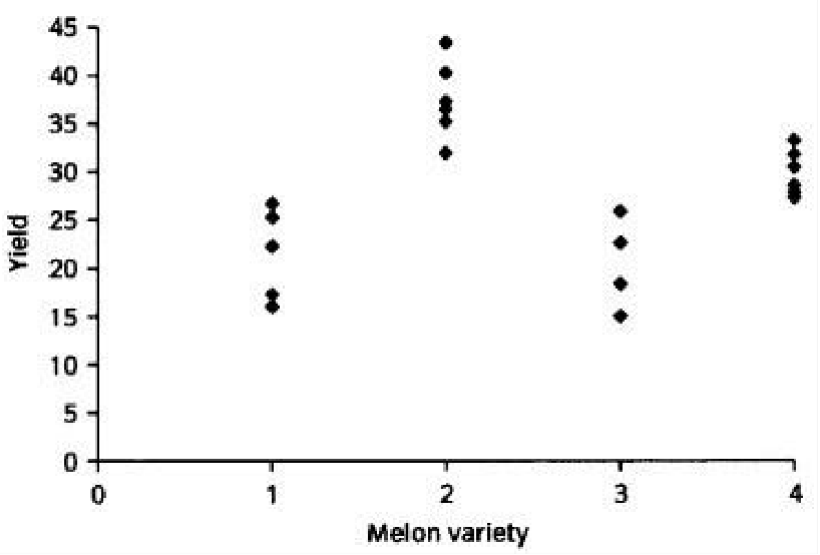

Fig. 1.15 Melon yields.

Table 1.7 ANOVA for *melons*

VARIETY	N	Mean
1	6	20.490
2	6	37.403
3	4	20.462
4	6	29.897

One-way analysis of variance for YIELDM

Source	DF	SS	MS	F	P
VARIETY	3	1115.3	371.8	23.80	0.000
Error	18	281.2	15.6		
Total	21	1396.5			

Dioecious trees

A plant species is dioecious if each individual produces all male flowers or all female flowers. The dataset *dioecious trees* contains data from 50 trees of one particular dioecious species, from a ten hectare area of mixed woodland. For each individual, the SEX was recorded (coded as 1 for male and 2 for female), the diameter at breast height in millimetres (DBH), and the number of flowers on the tree at the time of measurement (FLOWERS). This dataset will be revisited several times over the following chapters.

(1) Test the null hypothesis that male and female trees produce the same number of flowers.

(2) Show graphically how the number of flowers differs between the sexes.

Technical guidance on the analysis of these datasets is provided in the package specific supplements. Answers are presented at the end of this book.

2. Exercises

Does weight mean fat?

Can the weight of a person be used to predict how much body fat they are carrying around? In this study, total body fat was estimated as a percentage of body weight by using skinfold measurements of 19 students in a physical fitness program (stored in the dataset *reduced fats*). Weight was measured in kg.

Box 2.7 shows a regression analysis of these data and Fig. 2.16 a plot of these data.

(1) What is the best fitting straight line?

(2) What proportion of the variability in the data has been explained by fitting this line?

(3) How would you summarise the information provided by the data about the estimate of the slope?

(4) How strong is the evidence that the slope is different from zero?

(5) What would a zero slope imply about the relationship between the two variables?

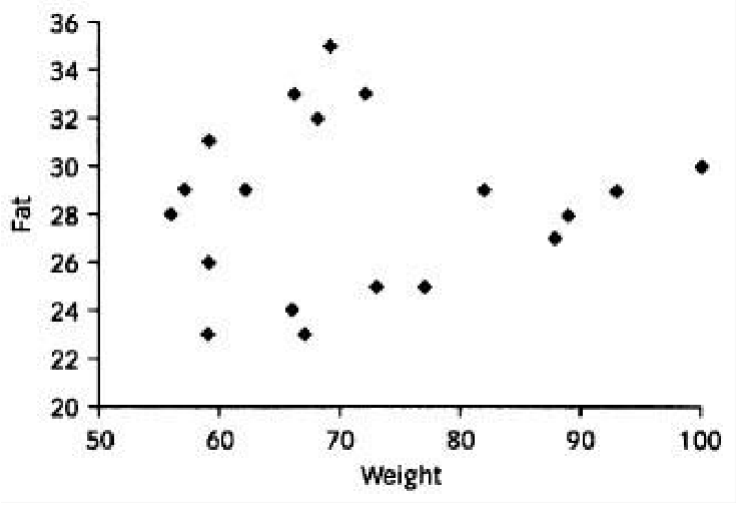

Fig. 2.16 Graph of *reduced fat* data.

BOX 2.7 Analysis of *reduced fat* data

Regression analysis

Word equation: FAT = WEIGHT

WEIGHT is continuous.

Analysis of variance table for FAT

Source	DF	SS	MS	F	P
Regression	1	1.33	1.33	0.10	0.751
Error	17	217.09	12.77		
Total	18	218.42			

Coefficients table

Predictor	Coef	SECoef	T	P
Constant	26.886	4.670	5.76	0.000
WEIGHT	0.02069	0.06414	0.32	0.751

Dioecious trees

This question returns to the *dioecious trees* datasct first used at the end of Chapter 1. The dataset contains three columns: FLOWERS, SEX and DBH (diameter at breast height).

(1) Illustrate graphically how FLOWERS and DBH are related.

(2) Using regression analysis, find the best fitting straight line predicting FLOWERS from DBH.

(3) Test the null hypothesis that the slope of the best fitting line equals 4.

4. Exercises

The cost of reproduction

Life history theory assumes that there is a trade off between survival and reproduction. Data were collected to test this assumption using the fruit fly *Drosophila subobscura.* Twenty-six female flies laid eggs over more than one day. Reproductive effort was measured as the average number of eggs laid per day over the lifetime of the fly. Survival was recorded as the number of days the fly survived after the first egg laying day. Their size was measured as the length of the prepupa at the beginning of the experiment, before emergence and egg laying began. Three variables were created: LSIZE, LLONGVTY and LEGGRATE, in which these data were logged. These variables are stored in the *Drosophila* dataset.

In Box 4.11 the researcher asked the question 'How does reproductive effort affect survival?'

BOX 4.11 GLM of survival against reproductive rate for *Drosophila subobscura*

General Linear Model

Word equation: LLONGVTY = LEGGRATE

LEGGRATE is continuous

Analysis of variance table for LLONGVTY, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
LEGGRATE	1	7.738	7.738	7.738	5.83	0.024
Error	23	30.507	30.507	1.326		
Total	24	38.245				

Coefficients table

Term	Coef	SECoef	T	P
Constant	1.7693	0.2313	7.65	0.000
LEGGRATE	0.2813	0.1165	2.42	0.024

BOX 4.12 GLM of survival against size and reproductive rate for *Drosophila subobscura*

General Linear Model

Word equation: LLONGVTY = LSIZE + LEGGRATE

LSIZE and LEGGRATE are continuous

Analysis of variance table for LLONGVTY, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
LSIZE	1	26.240	21.842	21.842	55.46	0.000
LEGGRATE	1	3.340	3.340	3.340	8.48	0.008
Error	22	8.665	8.665	0.394		
Total	24	38.245				

Coefficients table

Term	Coef	SECoef	T	P
Constant	1.6819	0.1266	13.28	0.000
LSIZE	1.4719	0.1976	7.45	0.000
LEGGRATE	-0.28993	0.09956	-2.91	0.008

A second analysis was then conducted in Box 4.12, which included the size of the flies.

(1) Calculate a confidence interval for the slope of LLONGVTY on LEGGRATE based on the analysis in Box 4.11.

(2) Calculate a confidence interval for the slope of LLONGVTY on LEGGRATE based on the analysis in Box 4.12, in which LSIZE has been eliminated.

(3) The graph of Fig. 4.10 is a plot of LLONGVTY against LEGGRATE with each point being allocated to one of six groups depending upon size (group 1 being the smallest up to group 6 being the largest). Why is there such a discrepancy between the two slopes at the centre of the confidence intervals calculated in 1 and 2?

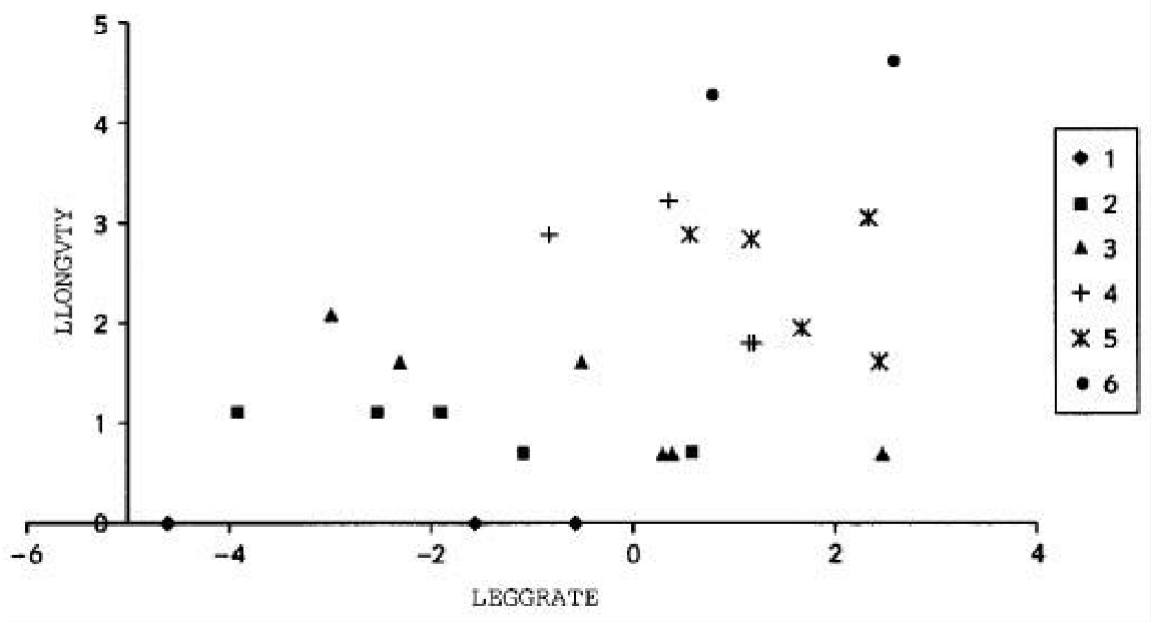

Fig. 4.10 LLONGVTY against LEGGRATE with each point being allocated to one of six size groups.

Investigating obesity

As part of an investigation into obesity, three measurements were taken from a sample of 39 men. These were: FOREARM, the thickness of a skin fold on the forearm, which is taken as an indicator of obesity; height (HT) and weight (WT). These data are found in the dataset *obesity*.

(1) Taking FOREARM as the response variable, which of the two explanatory variables HT and WT is the best predictor of obesity when used alone in a GLM?

(2) If the two explanatory variables are used together to predict FOREARM in a GLM, do they increase or detract from each other's informativeness and why?

(3) How could you predict the patterns found in the analyses you conducted in question 1 from the analysis you conducted in question 2?

5. Exercises

Growing carnations

A flower grower decided to investigate the effects of watering and the amount of shade on the number of saleable carnation blooms produced in his nursery. He designed his experiment to have three levels of watering (once, twice or three times a week) and four levels of shade (none, ¼, ½ and fully shaded). To conduct this experiment he needed to grow the carnation plots in three different beds. In case these beds differed in fertility or other important features, he decided to use these beds as blocks. After four weeks, he analysed the data by counting the number of blooms, and using the square root as the response variable, SQBLOOMS. He then fitted a GLM with three categorical explanatory variables: BED, WATER and SHADE (the data are stored in the *blooms* dataset). The output is shown in Box 5.6.

BOX 5.6 Analysis of the number of carnation blooms with bed, water and shade

General Linear Model

Word equation: SQBLOOMS = BED + WATER + SHADE

BED, WATER and SHADE are categorical variables

Analysis of variance table for SQBLOOMS, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
BED	2	4.1323	4.1323	2.0661	9.46	0.001
WATER	2	3.7153	3.7153	1.8577	8.50	0.001
SHADE	3	1.6465	1.6465	0.5488	2.51	0.079
Error	28	6.1173	6.1173	0.2185		
Total	35	15.6114				

Term	Coef	SECoef	T	P
Constant	4.02903	0.07790	51.72	0.000
BED				
1	0.0620	0.1102	0.56	0.578
2	0.3805	0.1102	3.45	0.002
3	*-0.4425*			
WATER				
1	-0.4110	0.1102	-3.73	0.001
2	0.3731	0.1102	3.39	0.002
3	*0.0379*			

SHADE				
1	0.0965	0.1349	0.72	0.480
2	0.2934	0.1349	2.17	0.038
3	-0.1191	0.1349	-0.88	0.385
4	*-0.2708*			

BOX 5.7 The carnation bloom analysis without bed used as a block

General Linear Model

Word equation: QBLOOMS = WATER + SHADE

WATER and SHADE are categorical variables

Analysis of variance table for SQBLOOMS, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
WATER	2	3.7153	3.7153	1.8577	5.44	0.010
SHADE	3	1.6465	1.6465	0.5488	1.61	0.209
Error	30	10.2496	10.2496	0.3417		
Total	35	15.6114				

Term	Coef	SECoef	T	P
Constant	4.02903	0.09742	41.36	0.000
WATER				
1	-0.4110	0.1378	-2.98	0.006
2	0.3731	0.1378	2.71	0.011
3	*0.0379*			
SHADE				
1	0.0965	0.1687	0.57	0.572
2	0.2934	0.1687	1.74	0.092
3	-0.1191	0.1687	-0.71	0.486
4	*-0.2708*			

1. Is the analysis in Box 5.6 orthogonal?

He then wondered whether it had been worth treating the beds as blocks, or whether future experiments could be fully randomised across beds. So he repeated the analysis without using bed as a blocking factor. This is shown in Box 5.7.

2. Was it worthwhile blocking for bed? If so, why?

He then discovered that a visitor had picked carnations from three of his experimental plots. He decided that because the final bloom numbers for these plots were inaccurate, he would exclude them from his analysis. So he produced a new, shorter data variable SQ2 for the square root of blooms, and explanatory variables B2, w2 and S2 for bed, water and shade levels respectively. This third analysis is presented in Box 5.8.

3. Which parts of the output differ in Box 5.8 but are the same in Box 5.6 and why? Does this fundamentally alter our conclusions?

4. Using the coefficient table given in Box 5.6, draw histograms illustrating how the number of blooms vary with level of water and level of shade.

BOX 5.8 Analysis of the carnation blooms with three plot values removed

General Linear Model

Word equation: SQ2 = B2 +W2 + S2

B2, W2 and S2 are categorical

Analysis of variance table for SQ2, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
B2	2	2.7626	2.6490	1.3245	8.03	0.002
W2	2	5.0793	4.6764	2.3382	14.18	0.000
S2	3	0.8072	0.8072	0.2691	1.63	0.207
Error	25	4.1213	4.1213	0.1649		
Total	32	12.7704				

The dorsal crest of the male smooth newt

Male smooth newts (*Triturus vulgaris*) develop a dorsal crest during the breeding season. During courtship the male releases pheromones and waggles his tail. The crest is thought to help the male waft the pheromones past the female's snout. A student conducted a survey to investigate variation in the size of the dorsal crest. She visited 10 local ponds, and measured a total of 87 male newts over a period of two weeks. The following data are recorded in the *newt* dataset:

LSVL: Logarithm of the snout-vent length in mm–a measure of skeletal size.

LCREST: Logarithm of the height of the dorsal crest in mm.

POND: A code from 1 to 10 for the pond at which the male was captured.

DATE: The day of the study on which the male was measured.

1. Taking LCREST as the response variable, analyse the data to investigate if the size of the dorsal crest reflects the body size of the newt.

2. Why is it probably a good idea to include POND in a model of this sort? Does it seem to matter in this case?

3. What circumstances might make it desirable to include DATE? How would you detect these circumstances?

6. Exercises

Conservation and its influence on biomass

An ecological study was conducted into the effect of conservation on the biomass of vegetation supported by an area of land. Fifty plots of land, each one hectare, were sampled at random from a ten thousand hectare area in Northern England. For each plot, the following variables were recorded:

BIOMASS: An estimate of the biomass of vegetation in kg per square metre.

ALT: The mean altitude of the plot in metres above sea level.

CONS: A categorical variable, which was coded as 1 if the plot was part of a conservation area, and 2 otherwise.

SOIL: A categorical variable crudely classifying soil type as 1 for chalk, 2 for clay and 3 for loam.

These data are stored in the *conservation* dataset. The output in Box 6.7 analyses BIOMASS as explained by the other three variables.

BOX 6.7 Conservation and biomass analysis

General Linear Model

Word equation: BIOMASS = CONS + ALT + SOIL

ALT is continuous, CONS and SOIL are categorical

Analysis of variance table for BIOMASS, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
CONS	1	0.7176	0.0249	0.0249	2.80	0.101
ALT	1	5.8793	4.4273	4.4273	498.10	0.000
SOIL	2	0.3953	0.3953	0.1977	22.24	0.000
Error	45	0.4000	0.4000	0.0089		
Total	49	7.3922				

Term	Coef	SECoef	T	P
Constant	2.21156	0.02486	88.97	0.000
CONS				
1	-0.02443	0.01460	-1.67	0.101
2	*0.02443*			
ALT	-0.002907	0.000130	-22.32	0.000
SOIL				

1	0.10574	0.02057	5.14	0.000
2	0.01952	0.01889	1.03	0.307
3	*-0.12526*			

1. On the basis of this analysis, what biomass would you predict for a plot with a mean altitude of 200m in a conservation area with chalk soil?

2. What biomass would you predict for a plot with mean altitude of 300m, with loam soil and not in a conservation area?

3. How strong is the evidence that the biomass of vegetation depends upon being in a conservation area? In what direction is the effect?

4. How strong is the evidence that SOIL affects BIOMASS? Which soil types are associated with the highest and lowest biomass values?

5. Give a 95% confidence interval for the effect of an additional metre of altitude on biomass.

6. Comment on the discrepancy between the sequential and adjusted sums of squares for CONS.

7. Given that it is impractical to conduct a randomised experiment in a study of this kind, what kind of uncertainties must remain in the conclusions that can be drawn?

Determinants of the Grade Point Average

The academic performance of some students in the USA is evaluated as a Grade Point Average (GPA) each year. Faculty are concerned to admit good students, and assess students via tests that are broken down into verbal skills (VERBA) and mathematical skills (MATH). A hundred students from each of two years (YEAR) had their marks analysed, to investigate whether verbal or mathematical skills were more important in determining a student's GPA. The variables GPA, YEAR, VERBAL and MATH are recorded in the *grades* dataset.

1. How good is the evidence that MATH, VERBAL or YEAR predicts GPA? In what direction is the effect for each of these variables?

2. What GPA would you expect from a student in the first year whose verbal score was 700 and mathematical score was 600? What about a student in the second year whose verbal score was 600 and mathematical score was 700?

7. Exercises

Antidotes

An experiment was conducted into the effectiveness of two antidotes to four different doses of toxin. The antidote was given five minutes after the toxin, and twenty-five minutes later the response was measured as the concentration of related products in the blood. There were three subjects at each combination of the antidote and dose level. The data are stored in the dataset *antidotes.* The results of the factorial ANOVA are given in Box 7.11.

(1) Draw the full interaction diagram.

(2) What are the conclusions from the ANOVA table?

(3) What is the most useful way to summarise the results of this experiment?

BOX 7.11 Interaction in the *antidotes* data

General Linear Model

Word equation: BLOOD = ANTIDOTE + DOSE + ANTIDOTE * DOSE

ANTIDOTE and dose are categorical

Analysis of variance table for BLOOD, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
ANTIDOTE	1	1396.90	1396.90	1396.90	23.68	0.000
DOSE	3	1070.09	1070.09	356.70	6.05	0.006
ANTIDOTE * DOSE	3	835.88	835.88	278.63	4.72	0.015
Error	16	943.68	943.68	58.98		
Total	23	4246.55				

Coefficients table

Term	Coef	SE Coef	T	P	
Constant	8.697	1.568	5.55	0.000	
ANTIDOTE					
1	7.629	1.568	4.87	0.000	
2	*-7.629*				
DOSE					
5	-8.186	2.715	-3.01	0.008	
10	-4.119	2.715	-1.52	0.149	
15	3.097	2.715	1.14	0.271	
20	*9.208*				
ANTIDOTE * DOSE					
1	5	-7.244	2.715	-2.67	0.017
1	10	-3.551	2.715	-1.31	0.209
1	15	2.573	2.715	0.95	0.358
1	*20*	*8.222*			
2	*5*	*7.244*			
2	*10*	*3.551*			

2	*15*	*-2.573*
2	*20*	*-8.222*

Weight, fat and sex

Returning to the *fats* data set, we are now in the position to do a more detailed analysis taking the sex of the participants into account. Analyse the data to answer the following:

(1) What is the best fitting line through the male data?

(2) What is the best fitting line through the female data?

(3) How strong is the evidence that the slopes differ?

15. Answers to exercises

Chapter 1

Melons

(1) The null hypothesis is that there is no difference in the mean yield between the varieties of melon.

(2) The null hypothesis is rejected (with $p < 0.0005$). We would conclude that there are significant differences in the mean yield of melons between the varieties. We estimate that variety 2 has the highest mean yield, and varieties 1 and 3 the lowest mean yields.

(3) The model produces an estimate of 15.6 for the unexplained variance with 18 degrees of freedom.

(4) The standard error of the mean is calculated by

$$\frac{s}{\sqrt{n}}$$

where

$$s = \sqrt{15.6} = 3.95.$$

This gives a standard error of 1.612 for varieties 1, 2 and 4, and a standard error of 1.975 for Variety 3.

(5) This information could be presented as means and their associated confidence intervals. The formula for a confidence interval is:

Mean $\pm$ $t_{crit}SE_{mean}$.

In this case, the critical t value could be for a 95% confidence interval and must have 18 degrees of freedom, giving 2.10. This gives the intervals presented in Table 15.1.

Table 15.1 Confidence intervals

Mean	95% Confidence interval
20.49	(17.11, 23.88)
37.40	(34.02, 40.79)
20.46	(16.32, 24.61)
29.90	(26.51, 33.28)

BOX 15.1 Analysis for dioecious trees

Word Equation: FLOWERS = SEX

SEX is categorical

Analysis of variance table for FLOWERS

Source	DF	SS	MS	F	P
SEX	1	171841	171841	1.18	0.284
Error	48	7017255	146193		
Total	49	7189097			

Dioecious trees

(1) sex is a categorical variable with two levels. To test the null hypothesis that male and female trees produce the same flowers, we need to fit the word equation

FLOWERS = SEX.

This would give the ANOVA table of Box 15.1.

We would therefore conclude that male and female trees do not have significantly different numbers of flowers.

(2) The data could be illustrated graphically in a number of ways. Here is a boxplot, in which the rectangle represents the middle 50% of the data, the line across the box being the median (middle value), and the tails stretching between the upper quartile and the maximum value, and between the lower quartile and the minimum. See Fig. 15.1.

This illustrates that while the medians are very close, female trees have much greater variability in the number of flowers than males.

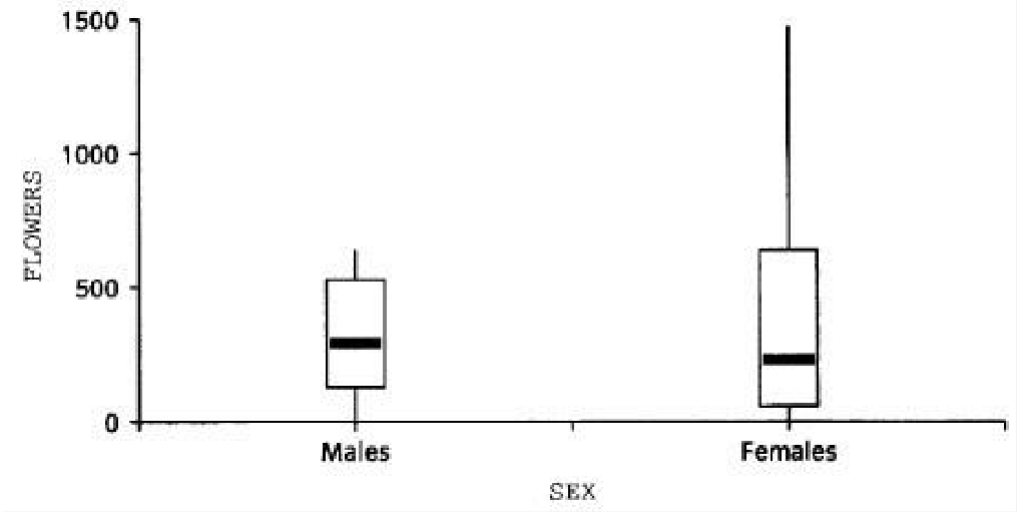

Fig. 15.1 Box plot for dioecious trees.

Chapter 2

Does weight mean fat?

(1) The fitted model is given by the equation FAT = 26.9 + 0.0207 WEIGHT.

(2) The proportion of explained variance is

$$\frac{1.33}{218.42} = 0.006.$$

(3) The slope is estimated to be 0.02069, with the standard error of this estimate as 0.06414. This information could also be presented as a confidence interval, with the critical *t* value having 17 degrees of freedom. This would give (-0.115, +0.156) as the 95% confidence interval for the slope.

(4) The slope is not significantly different from zero ($p = 0.751$).

(5) A zero slope would imply that WEIGHT provides no information about FAT.

Dioecious trees

(1) Plotting flowers against dbh gives the graph of Fig. 15.2.

(2) A regression analysis would use the word equation, FLOWERS = DBH, and would provide the output shown in Box 15.2.

This gives the fitted values equation as: FLOWERS = -481.16 + 4.5128 dbh.

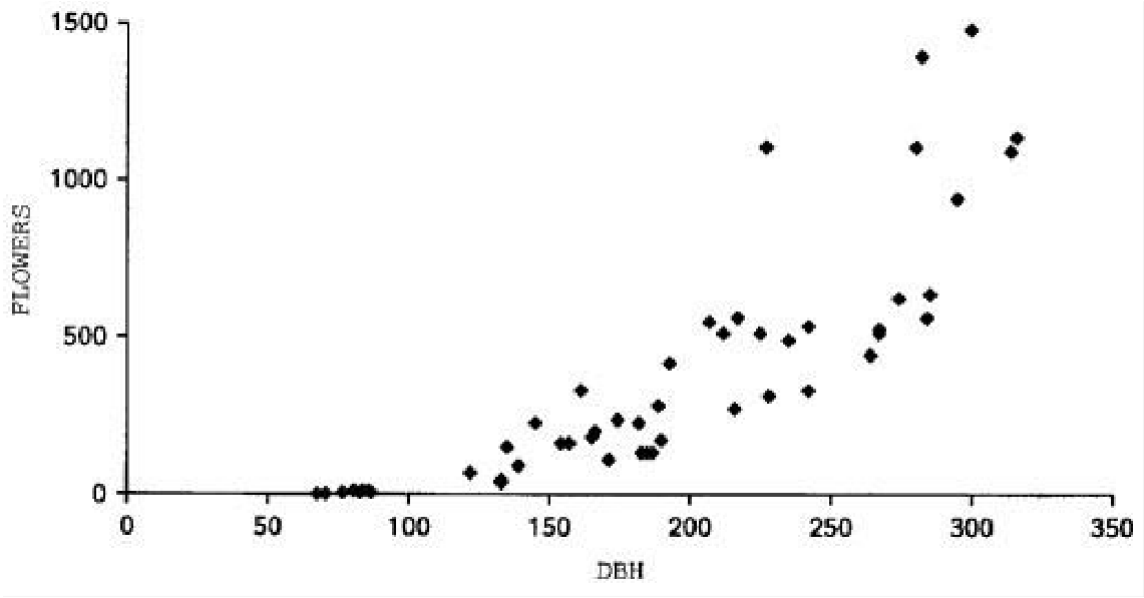

Fig. 15.2 Graph of FLOWERS versus DBH.

BOX 15.2 Analysis for dioecious trees

Regression analysis

Word Equation: FLOWERS = DBH

DBH is continuous

Analysis of variance table for FLOWERS

Source	DF	SS	MS	F	P
Regression	1	5060723	5060723	114.13	0.000
Residual Error	48	2128374	44341		
Total	49	7189097			

Coefficients table

Predictor	Coef	SECoef	T	P
Constant	-481.16	86.24	-5.58	0.000
DBH	4.5128	0.4224	10.68	0.000

(3) To test the null hypothesis that the slope is not significantly different from 4, we would calculate the test statistic as:

$$t_s = \frac{4.5128 - 4}{0.4224} = 1.21$$

for which $p = 0.232$. Therefore we conclude that the slope is not significantly different from 4.

Chapter 4

The cost of reproduction

(1) The 95% confidence interval for the slope of LLONGVTY on LEGGRATE is

$0.2813 \pm t_{23}.0.1165$

$= (0.0403, 0.5223)$

(2) The 95% confidence interval for the slope of LLONGVTY on LEGGRATE when size has been eliminated is

$-0.2899 \pm t_{22}.0.0996$

$= -0.2899 \pm 0.2066$

$= (-0.4965, -0.0833)$

(3) If you ignore the six different size categories, there is a positive relationship between survival and reproductive effort. However, it can also be seen that size is a confounding variable because the larger the flies are, the longer they live, and the more eggs they lay. Once the influence of size on reproduction and survival is eliminated, and we compare flies of the same size category, there is actually a negative relationship between reproductive effort and survival.

Investigating obesity

(1) Box 15.4 gives two separate analyses to explain FOREARM using HT or WT. From these analyses it would appear that WT alone is a better predictor of FOREARM than HT alone, giving a sum of squares of 59.137 compared to 0.944. That someone's weight can act as a predictor of obesity, but their height cannot, seems intuitively sensible.

(2) Box 15.5 shows the analysis using both explanatory variables together. From this it can be seen that the *F*-ratios (based on the adjusted sums of squares) for both WT and HT have increased–so together they increase each other's informativeness. In fact, HT is now significant ($p = 0.009$). This is because the combination of someone's height and weight provides much better predictive power for obesity than knowing one or other of these pieces of information.

BOX 15.4(a) First analysis of FOREARM

General Linear Model

Word equation: FOREARM = HT

HT is continuous

Analysis of variance table for FOREARM, using Adjusted SS for tests

Source	DF	Seq SS	AdjSS	AdjMS	F	P
HT	1	0.944	0.944	0.944	0.18	0.678
Error	37	199.094	199.094	5.381		
Total	38	200.038				

squares for height indicates that alone in the model it has poor explanatory power. The high

BOX 15.4(b) Second analysis of FOREARM

General Linear Model

Word equation: FOREARM = WT

WT is continuous.

Analysis of variance table for FOREARM, using Adjusted SS for tests

Source	DF	Seq SS	AdjSS	Adj MS	F	P
WT	1	59.137	59.137	59.137	15.53	0.000
Error	37	140.901	140.901	3.808		
Total	38	200.038				

BOX 15.5 Analysis of FOREARM using both explanatory varaiables

General Linear Model

Word equation: FOREARM = HT + WT

HT and WT are continuous

Analysis of variance table for FOREARM, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
HT	1	0.944	24.777	24.777	7.68	0.009
WT	1	82.970	82.970	82.970	25.72	0.000
Error	*36*	116.124	116.124	3.226		
Total	38	200.038				

Term	Coef	SE Coef	T	P
Constant	17.452	8.274	2.11	0.042
HT	-0.17173	0.06196	-2.77	0.009
WT	0.23317	0.04598	5.07	0.000

(3) These patterns and conclusions could be predicted from the third analysis by comparing the sequential and adjusted sums of squares for the two explanatory variables. There is no difference for the variable WT, but a substantial difference for the variable HT. The low sequential sum of adjusted sum of squares however indicates improved explanatory power with WT in the model.

Chapter 5

Growing carnations

(1) Yes–the sequential and adjusted sums of squares are identical, indicating that this data set is orthogonal.

(2) Yes–it was worthwhile blocking for BED, as this explained a significant amount of variation. Without BED as an explanatory variable, the variation explained by BED has been left as error variation, so reducing the *F*-ratios for the two treatments (and reducing the precision of all parameter estimates).

(3) The sequential and adjusted sums of squares are no longer exactly the same in Box 5.8, owing to loss of orthogonality. However, the differences are only slight, and do not alter our conclusions about which variables are significant.

(4) The two graphs are displayed in Fig. 15.4.

The dorsal crest of the male smooth newt

(1) The analysis of Box 15.6 shows that LSVL is a significant predictor of lcrest giving $p < 0.0005$.

(2) It is a good idea to include POND as local conditions may well influence the relationship between LCREST and LSVL. In this particular study however POND is insignificant, so its inclusion does not matter in this case as $p = 0.881$, see Box 15.7.

(3) If the data were collected over the breeding season, it may well have been the case that the crest was growing during the course of the study–and

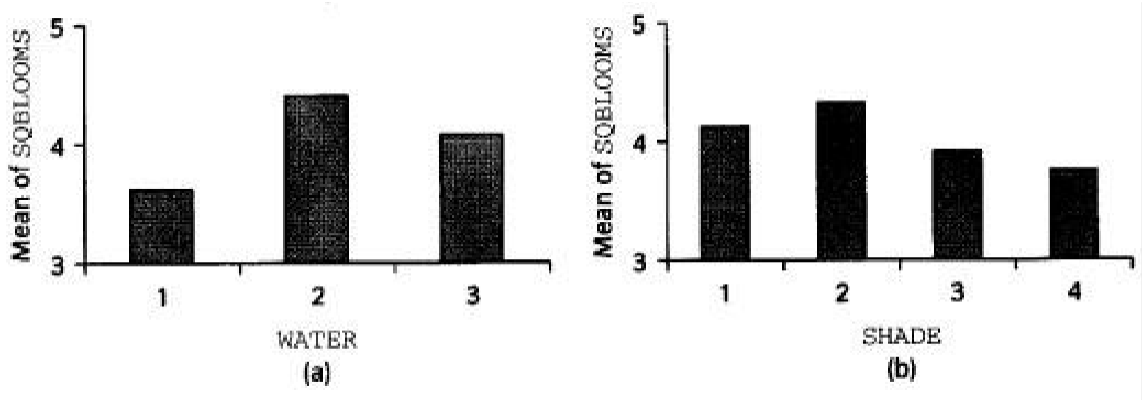

Fig. 15.4 Barcharts for SQBLOOMS.

changes. To detect these circumstances, DATE could be included as a continuous variable in the

BOX 15.6 Dorsal crest analysis

General Linear Model

Word equation: LCREST = LSVL

LSVL is continuous

Analysis of variance table for LCREST, using Adjusted SS for tests

Source	DF	SeqSS	AdjSS	AdjMS	F	P
LSVL	1	2.3894	2.3894	2.3894	45.81	0.000
Error	85	4.4337	4.4337	0.0522		
Total	86	6.8231				

Coefficients table

Term	Coef	SECoef	T	P
Constant	-9.381	1.501	-6.25	0.000
LSVL	5.0870	0.7516	6.77	0.000

BOX 15.7 Further dorsal crest analysis

General Linear Model

Word equation: LCREST = POND + LSVL

POND is categorical, LSVL is continuous

Analysis of variance table for LCREST, using Adjusted SS for tests

Source	DF	SeqSS	AdjSS	AdjMS	F	P
POND	9	0.32519	0.24063	0.02674	0.48	0.881
LSVL	1	2.30483	2.30483	2.30483	41.78	0.000
Error	76	4.19310	4.19310	0.05517		
Total	86	6.82312				

therefore the relationship between LCREST and LSVL would almost certainly have changed as the data were collected. In this case, inclusion of DATE could eliminate any such seasonal effects, and allow the relationship between LCREST and LSVL to be investigated over and above any seasonal model: if significant then seasonal effects would be important.

Chapter 6

Conservation and its influence on biomass

(1) BIOMASS = 2.21156 - 0.02443 - 0.002907 × 200 + 0.10574 = 1.711 (to 3 decimal places).

(2) BIOMASS = 2.21156 + 0.02443 - 0.002907 × 300 - 0.12526 = 1.239 (to 3 decimal places).

(3) The evidence that the biomass of vegetation depends upon being in a conservation area is weak ($p = 0.101$, which is not significant). Biomass of vegetation in the sample was lower in conservation areas.

(4) There is strong evidence that soil type affects biomass ($p < 0.0005$), biomass being highest on chalk and lowest on loam.

(5) The slope of altitude on biomass represents the effect of an additional metre on biomass, a 95% confidence interval being given by:

$-0.002907 \pm t_{45} \times 0.00013$

$= -0.002907 \pm 2.0141 \times 0.00013$

$= (-0.00317, -0.00265)$.

(6) The adjusted sum of squares for CONS is considerably lower than the sequential. This is due to a sharing of information between CONS, ALT and SOIL. This suggests that conservation areas tend to be more common at particular altitudes and/or on particular soil types. The very strong effect of cons, based on Seq SS, shows that it has a high correlation with BIOMASS. However, the low Adj SS suggests that this could be accounted for by a correlation with ALT or SOIL.

(7) Randomised experiments allow us to infer causation from correlation, while observational studies do not.

Determinants of Grade Point Average

(1) An analysis of the *grades* data set is given in the Box 15.8. There is no evidence that either YEAR or MATH predict GPA($p = 0.094$ and 0.124 respectively). The scores in the sample were higher for Year 1, and the trend in the sample was for higher MATH scores to be associated with higher gpa–but this trend is very slight, so the evidence that this is the case for the population is weak. The evidence that VERBAL predicts GPA is strong ($p < 0.0005$), with higher VERBAL scores being associated with higher GPA.

(2) For the first year with a VERBAL score of 700 and a MATH score of 600: GPA = 0.6582 + 0.06521 + 700 × 0.002288 + 600 × 0.000937 = 2.887. For the second year with a VERBAL score of 600 and a MATH score of 700: GPA = 0.6582 + (-0.06521) + 600 × 0.002288 + 700 × 0.000937 = 2.622.

BOX 15.8 Analysis of the grades dataset

General Linear Model:

Word equation: GPA = YEAR + VERBAL + MATH

YEAR is categorical, VERBAL and MATH are continuous

Analysis of variance table for GPA, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
YEAR	1	1.1552	0.8460	0.8460	2.84	0.094
VERBAL	1	6.7595	5.1600	5.1600	17.32	0.000
MATH	1	0.7092	0.7092	0.7092	2.38	0.124
Error	196	58.3961	58.3961	0.2979		
Total	199	67.0200				

Coefficients table

Term	Coef	SE Coef	T	P
Constant	0.6582	0.4404	1.49	0.137
YEAR				
1	0.06521	0.03870	1.69	0.094
2	-0.06521			
VERBAL	0.002288	0.000550	4.16	0.000
MATH	0.000937	0.000608	1.54	0.124

Chapter 7

Antidotes

(1) The coefficients for the full model are given in Table 15.2. The coefficients give the interaction diagram of Fig. 15.5.

(2) From the ANOVA table, we conclude that the interaction ANTIDOTE × DOSE is significant ($p = 0.015$). In other words, the effectiveness of the two antidotes changed depending upon dose of toxin administered, to different degrees. The interaction diagram suggests that the effectiveness of Antidote 2 changed little with toxin dose, but with Antidote 1 there was a very marked change.

(3) Given that the interaction is significant, then the 'one complicated story' illustrated by the interaction diagram is a good way to present these results. An alternative would be to present the two by four table of means and their associated standard errors as in Table 15.3.

Table 15.2 Coefficients for antidote analysis

	DOSE			
ANTIDOTE	5	10	15	20
1	0.897	8.657	21.997	33.757
2	0.127	0.500	1.593	2.053

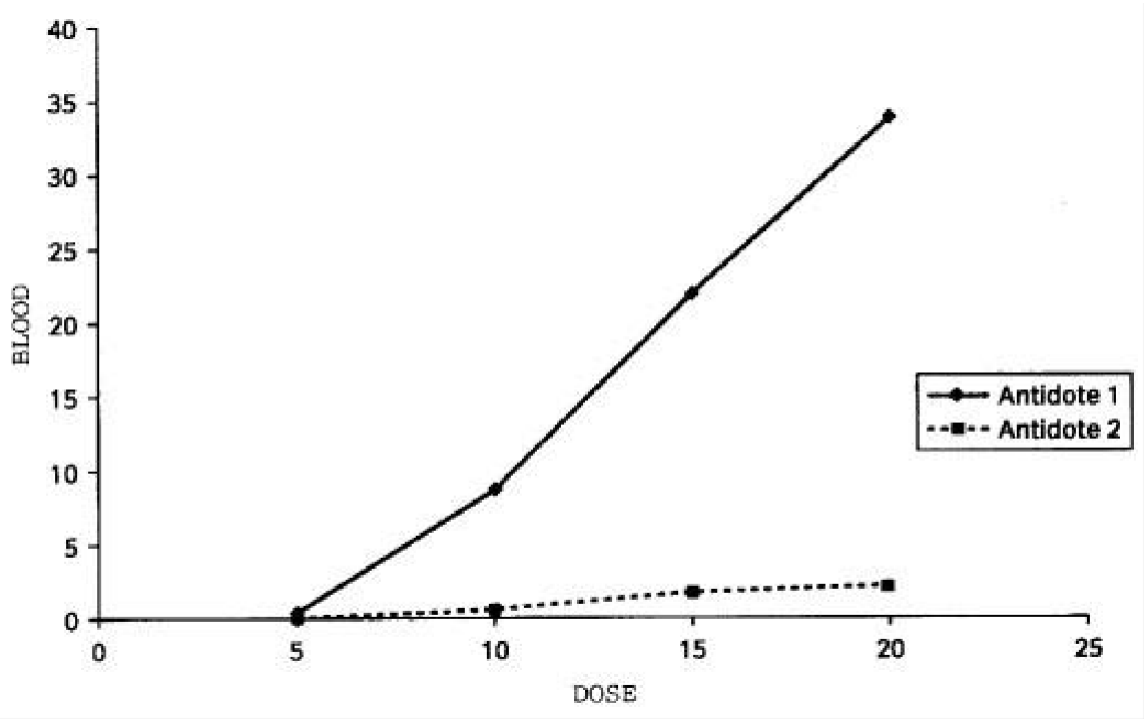

Fig. 15.5 Interaction diagram for antidote analysis.

Table 15.3 Means and their standard errors

ANTIDOTE × DOSE		Mean	SE Mean
1	5	0.8967	4.434
1	10	8.6567	4.434
1	15	21.9967	4.434
1	20	33.7567	4.434
2	5	0.1267	4.434
2	10	0.5000	4.434
2	15	1.5933	4.434
2	20	2.0533	4.434

There is another interesting point that we should notice about this table, and that is that the standard error of the mean is greater than the mean itself in many cases. The standard deviation (*s*) must be even greater. If *s* is a good estimate of the error standard deviation across the whole dataset, this implies negative concentrations, which are clearly nonsensical. The more obvious conclusion is that *s* is not a good estimate of error SD for all treatment combinations, and that the variance is heterogeneous. In Chapter 9 we will discuss how to deal with this.

Weight, fat and sex

The analysis in Box 15.9 fits the male and female relationships in the same analysis by use of the interaction term.

(1) This give the equation FAT = 11.571 + 0.1855 × WEIGHT for males.

(2) This give the equation FAT = 5.239 + 0.4029 × WEIGHT for females.

(3) There is evidence that the slopes differ, the interaction term having a *p*-value of 0.035.

BOX 15.9 Analysis for weight, fat and sex

General Linear Model

Word equation: FAT = SEX + WEIGHT + SEX * WEIGHT

SEX is categorical and WEIGHT is continuous

Analysis of variance table for fat, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
SEX	1	90.321	2.108	2.108	1.05	0.322
WEIGHT	1	87.105	79.542	79.542	39.59	0.000
SEX * WEIGHT	1	10.857	10.857	10.857	5.40	0.035
Error	15	30.138	30.138	2.009		
Total	18	218.421				

Coefficients table

Term	Coef	SE Coef	T	P
Constant	8.405	3.091	2.72	0.016
SEX				
1	-3.166	3.091	-1.02	0.322
2	3.166			
WEIGHT	0.29420	0.04676	6.29	0.000
WEIGHT * SEX				

1	0.10869	0.04676	2.32	0.035
2	-0.10869			