The Open University

Mathematics, Computing and Technology
M346 Linear Statistical Modelling

M346

Assignment Booklet

Contents **Cut-off date**

You will find instructions on how to fill in the PT3 form in the current *Assessment Handbook*. Remember to fill in the correct assignment number and to allow sufficient time in the post for each assignment to reach its destination on or before the cut-off date.

The marks allocated to each part of each question are indicated in brackets in the margin.

Many of these assignment questions require you to use your PC to analyse datafiles that were not installed during the main GenStat installation. These new files are supplied online through the M346 website, where instructions for their installation are also given.

WEB 03520 1
5.1

Questions 1 to 3 below, on *Units 1* to *3*, form Tutor-marked Assignment M346 01. Question 1 is marked out of 33; Question 2 is marked out of 35; Question 3 is marked out of 32. (The whole TMA is marked out of 100.)

The new datafiles used in this TMA and instructions for their installation (if you have not already installed them) are available from the M346 website.

You should be able to answer this question after you have studied Unit 1.

You do not need to use your computer to answer this question. You will, however, need to do some arithmetical calculations on your calculator.

Question 1 – 33 marks

In a study of 100 normal newborn babies, the level of the hormone 17-hydroxypregnenolone was measured in the umbilical blood. The observations for the first five babies, in nmol/l, are given in Table 1.

Table 1

17-hydroxypregnenolone
17.7
16.8
37.8
32.1
16.7

(a) A histogram of the data, with the fitted normal curve added, is given in Figure 1.

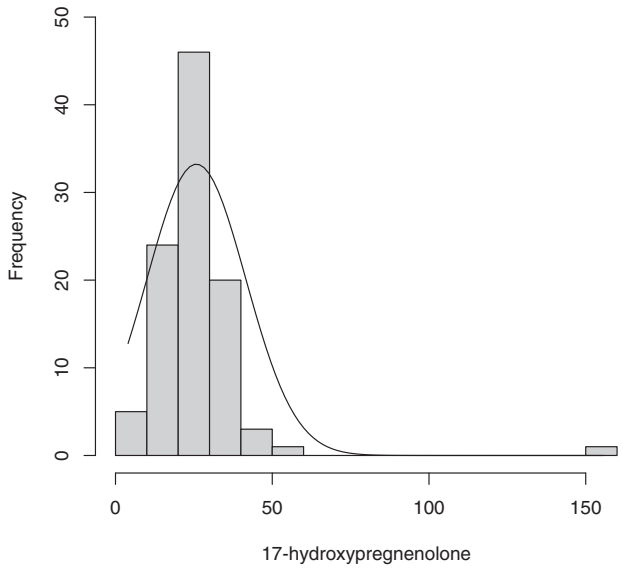


Figure 1

(i) Give two reasons why the data do not appear to be normally distributed. [*2*]

(ii) Give a reason why it is also not sensible to assume that the data follow a Poisson distribution. [*1*]

(b) Another possibility is that the data can be assumed to be normally distributed once the largest observation is dropped from the analysis. A normal probability plot based on the remaining 99 observations is given in Figure 2. Also given in Figure 2 are normal probability plots of three transformations of the data: x^{-1} (reciprocal of x), $\log(x)$ and x^2.

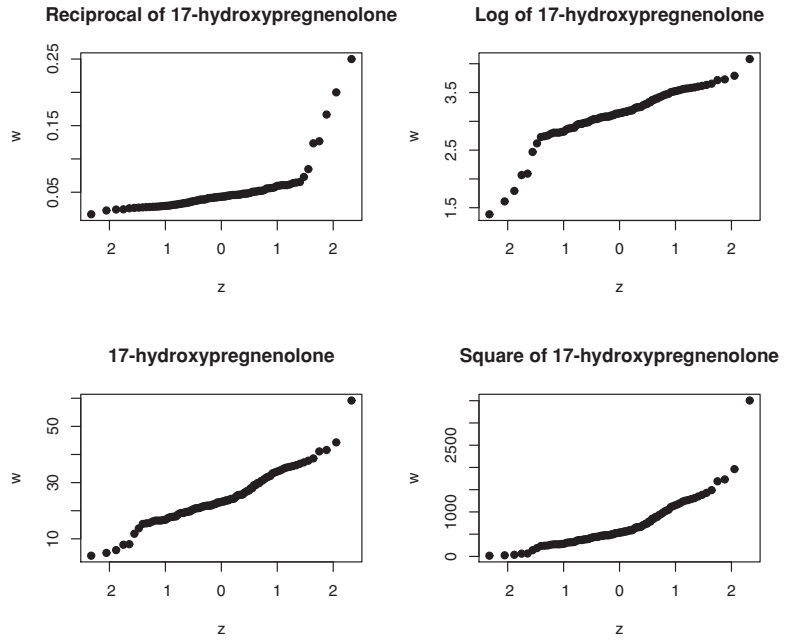


Figure 2

In which case does the assumption of normality appear to be the most plausible? Justify your choice. In this case, would you say that the data are represented well by a normal distribution? Why or why not? [*4*]

(c) Regardless of your answer to part (b), now assume that the (untransformed) data are normally distributed when the highest observation (155.6 nmol/l) is excluded from the dataset. For the remaining 99 observations, the mean 17-hydroxypregnenolone is 24.48 nmol/l, and the standard deviation is 8.79 nmol/l. Construct an (exact) 95% confidence interval for the mean 17-hydroxypregnenolone. Also construct a 95% confidence interval for the variance of the 17-hydroxypregnenolone levels. (Some potentially useful quantiles are given in Table 2.) [*8*]

Table 2

	Quantile							
Distribution	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
$t(98)$	−2.365	−1.984	−1.661	−1.290	1.290	1.661	1.984	2.365
$t(99)$	−2.365	−1.984	−1.660	−1.290	1.290	1.660	1.984	2.365
$t(100)$	−2.364	−1.984	−1.660	−1.290	1.290	1.660	1.984	2.364
$\chi^2(98)$	68.40	72.50	76.16	80.54	116.3	122.1	127.3	133.5
$\chi^2(99)$	69.23	73.36	77.05	81.45	117.4	123.2	128.4	134.6
$\chi^2(100)$	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8
Normal	−2.326	−1.960	−1.645	−1.282	1.282	1.645	1.960	2.326

(d) Calculate an approximate 95% confidence interval for the mean 17-hydroxypregnenolone using quantiles of a standard normal distribution. Why does this interval not depend on an assumption of normality for the distribution of 17-hydroxypregnenolone? [4]

(e) Compare the 95% confidence intervals for the mean that you calculated in parts (c) and (d). Hence comment on the assumption of normality made in part (c). [3]

(f) One explanation for the very high level of 17-hydroxypregnenolone observed in one of the 100 babies is that in fact the baby suffers from 'dry skin' disease (a condition in which 17-hydroxypregnenolone in umbilical blood is unusually high). It is estimated that 1 in every 4000 babies suffers from this disease.

What distribution can be used to model the number of babies in this study who suffer from 'dry skin' disease? State the parameters of the distribution in this case. (You may assume that the babies are all independent of each other.)

Hence calculate the mean number of babies who would be expected to suffer from 'dry skin' disease in a sample of 100 babies. In the light of this mean, comment on whether it is likely that the investigators had an affected baby in their sample. [4]

(g) Which of the following variables are nominal, which are ordinal and which are quantitative? For the quantitative variables, state whether they are discrete or continuous.

(i) Colour of newborn babies' eyes, recorded as brown, blue, green or other.

(ii) Amount of hair on a baby's head, recorded as none, a little, some, lots.

(iii) Birth weight, in grams.

(iv) Number of children the mother has (including the baby).

(v) Estimated length of the pregnancy (to the nearest week). [7]

You should be able to answer this question after you have studied Unit 2.

Question 2 – 35 marks

(a) Use GenStat to carry out the following probability calculations.

(i) Find $\Phi(1.6)$. [1]

(ii) Find $P(X > 5)$, where $X \sim \text{Poisson}(3)$. [1]

(iii) Find $P(-2.0 \leq T \leq 1.5)$, where $T \sim t(4)$. [1]

(iv) Find the 95% point of the $\chi^2(35)$ distribution. [1]

(v) Find $P(4 \leq Y \leq 7)$, where $Y \sim B(12, 0.4)$. [1]

(vi) Independent samples, of sizes $n_1 = 6$ and $n_2 = 8$, are taken from two populations that can be assumed to be normally distributed. The test statistic s_1^2/s_2^2 for a test that the two population variances are equal is calculated as 4.281. What is the two-sided p value for the test? [2]

(b) In a study of the growth of bark on cork oak, the weight of the cork deposit (in centigrams) was measured on 28 trees in each of the geographical directions north and east. The data are given in the GenStat datafile `cork.gsh`, with variates north and east. The observations on the first five trees are given in Table 3.

Table 3

north	east
72	66
60	53
56	57
41	29
32	32

[Source: Rao, C.R. (1948) 'Tests of significance in multivariate analysis', *Biometrika*, **35**, 58–79.]

(i) Calculate the differences between the cork deposit weights on the north and east sides of each tree. Using GenStat, draw (and send to your tutor) a histogram of the resulting differences using boundaries at $-15, -10, \ldots, 25$. Based on this, is it reasonable to assume that the differences are normally distributed? Why or why not? [*4*]

(ii) Whatever you concluded in part (b)(i), proceed as if the assumption of normality of the differences were justified. Find a 95% confidence interval for the mean difference in cork deposit weight between the north and east sides of a cork oak tree. Is it plausible that the cork deposit weight on cork trees does not depend on geographical direction? [*4*]

(iii) Use GenStat to draw a histogram of the cork deposit weights on the north sides of the cork trees. Send your histogram to your tutor. Is it reasonable to assume that the cork deposit weights on the north sides of cork trees have a normal distribution? Justify your answer. Briefly explain why it is irrelevant for the purposes of calculating the confidence interval in part (b)(ii) whether the distribution of cork deposit weights on the north sides is normal or not. [*3*]

(c) Ecologists collected data from a meadow at Tadham Moor in Somerset, England. For a large number of small areas (quadrats) in the meadow, they measured two so-called sum exceedence values that recorded, respectively, the extent to which each of the quadrats was subject to drought and to waterlogging. They also recorded the relative abundance of certain plant species in each of the quadrats. From these data, they calculated, for each of the plant species, two variables that measure the weighted averages of the susceptibilities to drought and to waterlogging of the parts of the meadow where the species grew. The resulting data for 21 species of the family *Poaceae* (grasses) and 7 species of the family *Cyperaceae* (sedges) are given in the datafile `meadow.gsh`. There are three columns. The factor family records the family of each of the 28 plant species included (with labels abbreviated to Po and Cy). The two variates, drought and waterlog, record the data on the average water conditions where each of the species grows; species that tend to grow in areas subject to drought have high values of drought, and those that tend to grow in areas subject to waterlogging have high values of waterlog.

[Source: Silvertown, J., Dodd, M.E., Gowing, D.J.G. and Mountford, J.O. (1999) 'Hydrologically defined niches reveal a basis for species richness in plant communities', *Nature*, **400**, 61–3.]

(i) In GenStat, produce boxplots (on the same diagram) of the values of drought for the two plant families. (To do this, in the Boxplot dialogue box, you will need to enter drought as the Data and family as the Groups.) Produce corresponding plots for the values of waterlog for the two families. Print all your plots to send to your tutor. On the basis of your boxplots, comment briefly on how the two plant families differ in terms of the average water conditions where the species grow. [*6*]

(ii) Produce the boxplots for drought again, but this time choose Variable for the Boxwidth in the Boxplot dialogue box. Print your plot to send to your tutor. Explain the difference between these boxplots and those that you produced for the same variable in part (c)(i). Use GenStat's Help facilities to find out precisely what the difference is. [*3*]

(iii) Use your boxplots to comment on whether assumptions of normality for the four sets of data seem justified. [*4*]

(iv) Whatever you concluded in part (c)(iii), proceed as if assumptions of normality are justified for the waterlog variate. Use GenStat to perform an appropriate (two-sided) t-test of the null hypothesis that the data on this variate for the two plant families are drawn from populations with equal means. (You will need to change the default Data Arrangement in the Two-sample Tests dialogue box.) What is the SP for this test? What conclusion do you reach? [*4*]

You should be able to answer this question after you have studied Unit 3.

Question 3 – 32 marks

The data in the GenStat datafile `value.gsh`, of which the first five datapoints are shown in Table 4, were collected some time ago by Kevin McConway, one of the original authors of M346. His aim in doing so was to try to use sampling methods to estimate the replacement cost of a large collection of books, for insurance purposes, on the basis of a sample of 100 of them. Replacement prices for the sample of books (price, in pence) were found from publishers' catalogues. An attempt was made to improve the accuracy of the estimated total replacement value by measuring another variable, the width (in mm) of the spine of the book, which (it was thought) might be related to the cost. In this question, you are asked to explore the relationship between price and width.

Table 4

price	width
995	24
1250	13
295	33
295	2
250	11

(a) Produce a scatterplot of the data, treating width as the explanatory variable. Comment on the relationship between the width of a book and its replacement value. [4]

(b) Fit a regression line and report its equation. Also produce a composite residual plot. Hence comment on the appropriateness of the model. [6]

(c) Supposing that there is indeed a problem with the regression assumptions about the error but not with the linearity of the mean relationship, give an argument for not transforming either of the variables singly.

In fact, an appropriate approach is to transform both variables simultaneously by the same transformation. Whereabouts on the ladder of powers might you expect to find an appropriate transformation? [3]

(d) Experiment with appropriate transformations of both variables, using the associated residual plots to identify a good transformation. You need only show a composite residual plot for the transformation that you prefer. [6]

(e) Although you may very well have argued in favour of some alternative transformation, work now with taking logs of each variable. What is the fitted regression line in this case? By taking exponentials of both sides of the fitted equation (remember that $e^{a+b\log x} = cx^b$ where $c = e^a$), re-express the model for price in terms of width: in round terms, to what power of width does price appear to be roughly proportional? [4]

(f) According to the model fitted in part (e), what is the estimated price of a book whose spine is 16.2 mm wide? Obtain a 95% prediction interval for the price of this book.

Given the interval that you obtain, if you were told that a 95% confidence interval for the mean price of a book whose spine is 16.2 mm wide is $(99.6, 2075.1)$, how would you know without any further calculations that this confidence interval was in error? [7]

(g) Professor McConway's complete collection of books numbered some 1554 volumes, with mean spine width 16.2 mm. Using your answer to part (e), what is the point estimate of the replacement value of his entire collection? [2]

[Congratulations: you have just re-invented a standard method for so-called finite population inference (if not quite in precise detail), a topic not covered in general in this module!]

Questions 1 to 3 below, on *Units 4* to *6*, form Tutor-marked Assignment M346 02. Question 1 is marked out of 32; Question 2 is marked out of 35; Question 3 is marked out of 33. (The whole TMA is marked out of 100.)

The new datafiles used in this TMA and instructions for their installation (if you have not already installed them) are available from the M346 website.

You should be able to answer this question after you have studied Unit 4.

Question 1 – 32 marks

The data stored in the file `mnemonic.gsh` were obtained in tests of ways to improve memory. Two mnemonic methods for trying to improve verbal recall are the Galton's walk method (Mnemonic A) and the peg method (Mnemonic B). Thirty participants were randomly assigned to one of three equal-sized groups:

- a group trained to use Mnemonic A (group A);
- a group trained to use Mnemonic B (group B);
- a control group, which received no training (group C).

Each participant was presented with verbal material, and at a later stage was asked to reproduce it in free written recall. In the datafile, the first column (count) gives the number of words recalled by a participant, and the second column (group) gives the participant's group.

[Source: Kinnear, P.R. and Gray, C.D. (1996) *SPSS for Windows Made Simple*, Hove, Psychology Press.]

(a) Is this study an observational study or a controlled experiment? Give a reason for your answer. [*3*]

(b) Use GenStat to produce an appropriate table of summary statistics and an appropriate graphical display to illustrate the number of words recalled for the different groups.

On the basis of your table and diagram, comment on whether and, if so, how the distribution of the number of words recalled differs between the three groups.

On the basis of your table and diagram, also comment on whether these data appear to satisfy the assumptions for analysis of variance. [*6*]

(c) Now, whatever your answer to part (b), assume that it is sensible to carry out an analysis of variance for these data. Use the GenStat analysis of variance commands to obtain the ANOVA table, and test the hypothesis that word recall is affected by training/type of training in a mnemonic technique. Include appropriate GenStat printout to support your conclusions. [*5*]

(d) Produce appropriate residual plots to check the appropriateness of the analysis of variance model fitted in part (c). Comment, in the light of the plots, on the adequacy of the model. [*5*]

(e) Denote by μ_A, μ_B and μ_C the expected number of words recalled by a person in groups A, B and C, respectively. Consider the contrasts

$$\theta_1 = \tfrac{1}{2}(\mu_A + \mu_B) - \mu_C \quad \text{and} \quad \theta_2 = \mu_A - \mu_B.$$

Using GenStat, expand the ANOVA table to test the hypothesis that $\theta_1 = 0$. Then expand the ANOVA table to test the hypothesis that $\theta_2 = 0$.

Give the p values for each hypothesis test, and the conclusions from the hypothesis tests. Interpret the results in terms of the tests to improve memory. [8]

(f) The memory test was repeated, but with 12 participants in each group. Part of the resulting ANOVA table is reproduced below. Showing your working, complete the table. [5]

Variate: count

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
group	XX	XXXXXXX	XXXXXX	XXXXX	XXXXX
Residual	XX	402.90	XXXXXX		
Total	35	1067.70			

You should be able to answer this question after you have studied Unit 5.

Question 2 – 35 marks

The dataset stored in the file `swiss.gsh` consists of a standardised fertility measurement (fertilty), I_g, for 47 French-speaking provinces of Switzerland in about 1888, together with the following socio-economic indicators for the same provinces:

agricltr	percentage of the population involved in agriculture as an occupation;
examintn	percentage of drafted soldiers receiving the highest mark on the army examination;
educatin	percentage of the population educated beyond primary school;
catholic	percentage of the population who were Catholic;
mortalty	percentage of live births who lived less than one year (this is called *infant mortality*).

[Source: Mosteller, F. and Tukey, J.W. (1977) *Data Analysis and Regression: A Second Course in Statistics*, Reading, MA, Addison-Wesley, pp. 549–51.]

The aim is to determine whether, and how, fertility is related to the five socio-economic variables.

(a) Using GenStat, produce a scatterplot matrix including all the variables, and also produce their correlation matrix. On the basis of these, carry out a preliminary examination of the relationships between fertilty and the explanatory variables, and of the relationships between explanatory variables. Regarding fertilty as the response, which of the other variables would you expect to see as explanatory variables in a good linear regression model? [8]

(b) (i) Fit a regression model including all five explanatory variables. In this model, which variables seem to be important? [2]

(ii) Produce a composite residual plot. Judging from the composite residual plot, do the assumptions seem to be reasonable? [3]

(iii) Your GenStat output should contain the following message.

Message: the following units have large standardized residuals.

Unit	Response	Residual
37	92.2	2.31
47	42.8	–2.27

What does this message mean? Why might this be important? Is this message a cause for concern in this particular case? [3]

(iv) Your GenStat output should also contain the following message.

Message: the following units have high leverage.

Unit	Response	Leverage
19	54.3	0.35
45	35.0	0.46

You are not expected to know what this message means yet; for now, accept that it means that there is something unusual about provinces 19 and 45.

On the scatterplot matrix that you produced in part (a), mark (by hand) where the points corresponding to provinces 19 and 45 are. What is unusual about provinces 19 and 45? Why might this be important? In your opinion, is province 45 a rural province, or is it a province containing a large town or city? [7]

(c) (i) Perform a stepwise regression starting from the full regression model, using 16 as the maximum number of steps and 4 as the text criterion. Which explanatory variables are selected by this procedure? Are these the variables that you expected to be selected when you carried out your initial data examination in parts (a) and (b)(i)? [5]

(ii) Does the stepwise regression starting from the null model lead to the same selected model? [3]

(iii) Summarise your findings, including giving the fitted model. [4]

You should be able to answer this question after you have studied Unit 6.

Question 3 – 33 marks

An experiment was carried out to investigate the effect of three factors on the survival of the bacterium *Salmonella typhimurium*. Three levels of sorbic acid, three pH (acidity) levels and six levels of water activity were used. The experiment was run once at each possible combination of factor levels. The response variable was the logarithm of the density of bacteria (per millilitre) seven days after treatment started. The data are shown in Table 5.

Table 5

Sorbic acid (parts per million)	pH	Water activity					
		0.78	0.82	0.86	0.90	0.94	0.98
0	5.0	4.20	4.52	5.01	6.14	6.25	8.33
	5.5	4.34	4.31	5.35	5.98	6.70	8.37
	6.0	4.31	4.85	5.06	5.87	6.65	8.19
100	5.0	4.18	4.18	4.29	5.78	6.51	7.59
	5.5	4.39	4.43	4.95	5.28	6.19	7.79
	6.0	4.13	4.29	4.85	5.01	6.52	7.64
200	5.0	4.15	4.37	4.79	5.43	6.43	7.19
	5.5	4.12	4.27	4.40	5.10	6.18	6.92
	6.0	3.93	4.26	4.41	5.20	6.33	7.14

[Source: Mead, R. (1988) *The Design of Experiments*, Cambridge, Cambridge University Press.]

The data from Table 5 are stored in the file `salmonel.gsh`, with the response variable labelled as response, and the three treatment factors as sorbic, pH and activity. Load the data into GenStat. Assume that it is acceptable to analyse them using analysis of variance.

(a) State how many factors are involved in this experiment, and how many levels each factor has. [*2*]

(b) In this experiment, one observation was made for each treatment. Thus there is no way of measuring the variability within treatments.

(i) Demonstrate this by comparing the degrees of freedom of the data and those of the main effects and their interactions. [*2*]

(ii) What kind of assumption is made to get round this limitation and analyse the data? [*1*]

(c) Using GenStat's analysis of variance commands, a model was fitted that included all three main effects and all three two-way interactions (model A). The GenStat output is as follows.

Model A

Analysis of variance

Variate: response

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
activity	5	81.56910	16.31382	470.20	<.001
pH	2	0.01385	0.00692	0.20	0.821
sorbic	2	2.75936	1.37968	39.77	<.001
activity.pH	10	0.45191	0.04519	1.30	0.294
activity.sorbic	10	1.31626	0.13163	3.79	0.005
pH.sorbic	4	0.22806	0.05702	1.64	0.203
Residual	20	0.69391	0.03470		
Total	53	87.03245			

Message: the following units have large residuals.

units	8	–0.237	s.e. 0.113
units	21	–0.273	s.e. 0.113
units	39	0.231	s.e. 0.113

From the results, what conclusion do you reach about the effect of the factors? [2]

(d) To examine the assumptions underlying the model, a composite residual plot was produced for model A, and this is given in Figure 3. Why does the composite residual plot suggest that the data ought to be transformed? [2]

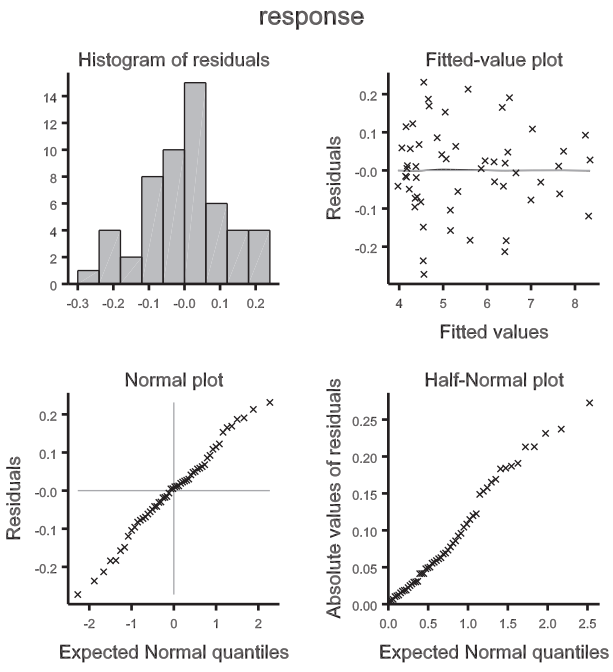


Figure 3

(e) For each of the three factors in turn, obtain the means and variances of the response variable, response, for each of its factor levels. Why does this suggest that the data ought to be transformed? [*4*]

(f) Calculate a new variable ires = 1/response. Using GenStat, fit an analysis of variance model with ires as the response, and include all three main effects and all the two-way interactions. From the results, what conclusion do you reach about the effect of the factors on ires? Produce and send to your tutor suitable residual plots. Do the assumptions of ANOVA appear reasonable in this case? Briefly compare your findings with model A. [*6*]

(g) Produce a means plot in which the effect of activity is along the horizontal axis, the groups are given by sorbic, and means corresponding to the same level of activity are joined by lines. Comment on how this plot reflects the p values associated with the activity, sorbic and activity*sorbic terms that you obtained in part (f). [*4*]

(h) Which of model A and the model that you fitted in part (f) do you consider to be best? Briefly explain your choice with respect to ANOVA assumptions and model parsimony. [*3*]

(i) An alternative approach to analyse these data would be to treat one or more of the explanatory variables as variates, rather than as factors, and to fit a regression model. Suggest, with a reason, one explanatory variable that might be a good candidate to convert to a variate. What would be the advantage of this approach over the ANOVA models fitted? [*4*]

(j) The output below is an analysis of variance table extracted from GenStat by fitting ires with an additive model involving the main effects of activity and sorbic only. Using the values that have been supplied, calculate the three numbers that should replace the crosses to complete the table. You are strongly advised to set your answers out clearly in the order in which you calculate them (making clear which number refers to which obscured value in the table), so that if you go wrong at an early stage, partial credit can still be given for the method that you use. [*3*]

Analysis of variance

Variate: ires

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
activity	5	XXXXXX	0.01615020	254.34	<.001
sorbic	2	0.00267411	0.00133705	XXXX	<.001
Residual	46	0.00292096	0.00006350		
Total	53	XXXXXX			

Questions 1 to 3 below, on *Units 7* to *9*, form Tutor-marked Assignment M346 03. Question 1 is marked out of 34; Question 2 is marked out of 33; Question 3 is marked out of 33. (The whole TMA is marked out of 100.)

The new datafiles used in this TMA and instructions for their installation (if you have not already installed them) are available from the M346 website.

You should be able to answer this question after you have studied Unit 7. You may also need to review Section 4.5 of Unit 4.

Question 1 – 34 marks

(a) Robert conducted an experiment to evaluate in which of five sound modes he best played a certain video game. The five sound modes were as follows:

- Sound modes 1, 2 and 3 corresponded to game sounds plus three different background music tracks.
- Sound mode 4 corresponded to game sounds, but no background music.
- Sound mode 5 corresponded to no game sounds and no background music.

Robert believed that his game performance, measured by game score, varied from day to day, and that he became tired or bored after 4 to 6 games. So he used a latin square design with two blocking factors: day, and time-order of game play. The data from this experiment are given in the file `videogames.GSH`. The response variable is labelled score, the sound modes are labelled sound, and the day and time-order are labelled day and time, respectively.

[Source: Dean A. and Voss D. (1999) *Design and Analysis of Experiments*, New York, Springer.]

(i) Produce a scatterplot of the game scores against sound mode, with day as the grouping factor. Then produce a second scatterplot of the game scores against sound mode, with time-order as the grouping factor. Include only one of these plots in your answer. Do there appear to be any differences between sound modes? Do either the day or the time-order appear to have any effect on the game score? [5]

(ii) Obtain the appropriate ANOVA table for this experiment, and use it to answer the question: 'Is there any difference in game performance between the five sound modes?' [4]

(iii) Robert constructed two contrasts as follows:

$$\theta_1 = \tfrac{1}{3}(\mu_1 + \mu_2 + \mu_3) - \tfrac{1}{2}(\mu_4 + \mu_5),$$
$$\theta_2 = \tfrac{1}{4}(\mu_1 + \mu_2 + \mu_3 + \mu_4) - \mu_5,$$

where μ_1 is Robert's mean score with sound mode 1, μ_2 is his mean score with sound mode 2, and so on.

Using the sound mode descriptions given at the beginning of the question, briefly describe the comparison made by θ_1 and θ_2.

Two new ANOVA tables were produced that each included one of these contrasts. The extra line in the ANOVA table with θ_1 was

contrast	d.f.	s.s.	m.s.	v.r.	F pr.
θ_1	1	319.7	319.7	2.19	0.164

and in the ANOVA table with θ_2 it was

contrast	d.f.	s.s.	m.s.	v.r.	F pr.
θ_2	1	492.8	492.8	3.38	0.091

Test the hypotheses $\theta_1 = 0$ and $\theta_2 = 0$. [*4*]

(iv) What do you conclude from this experiment? Which sound mode(s) should Robert use to optimise his game performance? [*2*]

(b) A food processing company wished to compare the taste of six new brands of breakfast cereal that were labelled A, B, C, D, E and F. 15 subjects were asked to taste four cereals and score them on a scale of 0–100. The data are given in Table 6.

Table 6

Subject	Cereal (A–F) and score (0–100)			
1	A 51	B 55	C 69	D 83
2	A 48	D 87	E 56	F 22
3	B 65	C 91	E 67	F 35
4	A 42	B 48	C 65	E 43
5	A 36	B 58	D 69	F 7
6	C 79	D 85	E 56	F 25
7	A 54	B 60	C 90	F 21
8	A 62	C 92	D 94	E 63
9	B 39	D 71	E 47	F 11
10	A 51	B 59	D 84	E 51
11	A 39	C 74	E 61	F 25
12	B 69	C 78	D 78	F 22
13	A 63	B 74	E 59	F 32
14	A 55	C 74	D 78	F 34
15	B 73	C 83	D 92	E 68

(i) Describe the features of this dataset that mean that the experiment follows a *balanced incomplete block design*, where each subject is treated as a block. [*5*]

(ii) The data are given in `cereal.GSH`, where the responses are labelled `score`. Compare the spreadsheet and Table 6, so as to ensure that you understand how the data in the spreadsheet correspond to the table. Obtain the ANOVA table for this experiment. What does it tell you about the differences in taste scores between cereals? [*5*]

(iii) To check the appropriateness of the model that GenStat is using, produce the usual set of residual plots. Are any of the assumptions of the model in doubt? [*4*]

(iv) The two top scoring cereal brands are D and C. Use the output from part (b)(ii) to calculate a point estimate of the difference in taste score between cereal D and cereal C, and to calculate a 95% confidence interval for this difference. Is it plausible that there is no difference in taste score between cereal brands C and D? [*5*]

You should be able to answer this question after you have studied Unit 8.

Question 2 – 33 marks

A test that is commonly used on samples of blood from human patients is to determine the erythrocyte sedimentation rate (ESR), which is the rate at which red blood cells (erythrocytes) settle out of suspension in blood plasma. The ESR tends to rise if the levels of certain proteins in the blood increase, and this happens in the presence of certain diseases (including some infectious, inflammatory and malignant diseases). The ESR may therefore be a useful diagnostic indicator for such diseases. A study was carried out by the Institute of Medical Research, Kuala Lumpur, in which (among other things) the ESR and the levels of two plasma proteins, fibrinogen (fibrinog) and γ-globulin (gglobuli), were measured in 32 individuals. The levels of the two plasma proteins are measured in grams per litre (g/l). The ESR is here recorded as a 0–1 indicator variable (esrind), with 0 denoting an ESR of less than 20 millimetres per hour (mm/h) (which is the usual observation in 'healthy' patients), and 1 denoting an ESR of 20 mm/h or more. (In all, six of the individuals studied have an ESR of 20 mm/h or more.)

[Source: Collett, D. and Jemain, A.A. (1985) 'Residuals, outliers and influential observations in regression analysis', *Sains Malaysiana*, **14**, 493–511.]

The aim of the analysis of these data is to investigate how (if at all) the probability of an ESR reading of 20 mm/h or more is affected by the levels of the two proteins.

(a) Figure 4 shows scatterplots of the binary response variable esrind against each of the explanatory variables fibrinog and gglobuli.

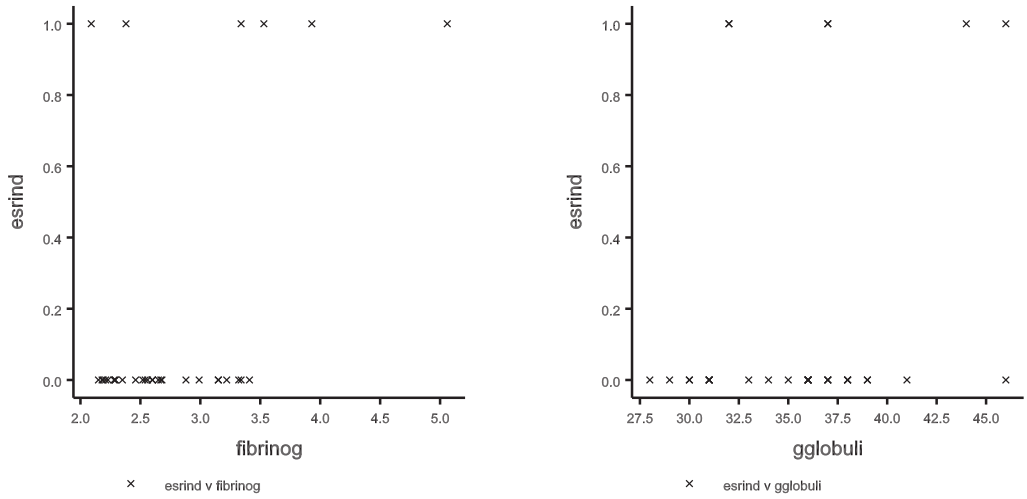


Figure 4

(i) On the basis of Figure 4, which of the two explanatory variables would you say has the stronger relationship with an ESR level of 20 mm/h or greater? Briefly explain your answer. [*2*]

(ii) Why is it difficult to interpret the plot for gglobuli in Figure 4? What other type of plot might it be useful to look at? [*2*]

(b) Would a logistic relationship between the explanatory variables and the response variable be appropriate for these data? Justify your answer. [*2*]

(c) Assume that a logistic regression model is appropriate for these data.

The following output is generated by GenStat from fitting logistic regression models to each of the two explanatory variables fibrinogen (Model A) and gglobuli (Model B). Only the summary of analysis table is given for Model B.

Model A

Regression analysis

Response variate:	esrind
Binomial totals:	1
Distribution:	Binomial
Link function:	Logit
Fitted terms:	Constant, fibrinog

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	1	6.04	6.0446	6.04	0.014
Residual	30	24.84	0.8280		
Total	31	30.88	0.9963		

Dispersion parameter is fixed at 1.00.

Message: deviance ratios are based on dispersion parameter with value 1.

Message: the following units have large standardized residuals.

Unit	Response	Residual
15	1.00	2.32
23	1.00	2.53

Message: the residuals do not appear to be random; for example, fitted values in the range 0.09 to 0.31 are consistently larger than observed values and fitted values in the range 0.40 to 0.92 are consistently smaller than observed values.

Message: the following units have high leverage.

Unit	Response	Leverage
13	1.00	0.276
29	1.00	0.226

Estimates of parameters

Parameter	estimate	s.e.	t(*)	t pr.	antilog of estimate
Constant	–6.85	2.76	–2.48	0.013	0.001065
fibrinog	1.827	0.899	2.03	0.042	6.216

Message: s.e.s are based on dispersion parameter with value 1.

Model B

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	1	1.94	1.9395	1.94	0.164
Residual	30	28.95	0.9648		
Total	31	30.88	0.9963		

Report the regression deviances and significance probabilities of each of models A and B, and discuss the implication for leaving out each of the two explanatory variables. [4]

(d) gglobuli was added to model A to give model C. The summary of analysis table for model C is given below.

Model C

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	2	7.91	3.9569	3.96	0.019
Residual	29	22.97	0.7921		
Total	31	30.88	0.9963		
Change	–1	–1.87	1.8692	1.87	0.172

What line of the output for model C allows for testing whether it is appropriate to leave gglobuli out of the model, in the presence of fibrinogen? What does this test show? Briefly explain. Which model would you select? [4]

(e) Denote by p the probability that a person's ESR is 20 mm/h or higher. Write down a formula giving the log odds corresponding to p, according to model A, when a person's plasma levels of fibrinogen is denoted by f. Also, write down a formula giving p directly. [4]

(f) The following GenStat output shows model D, which includes both fibrinog and its square transformation. (That is, model D fits the probability that the response variable takes the value 1 as a logistic function of a quadratic function of the explanatory variable fibrinog, rather than simply as a logistic function of a linear function on it.)

Model D

Regression analysis

Response variate:	esrind
Binomial totals:	1
Distribution:	Binomial
Link function:	Logit
Fitted terms:	Constant + fibrinog
Submodels:	POL(fibrinog; 2)

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	2	13.92	6.9587	6.96	<.001
Residual	29	16.97	0.5851		
Total	31	30.88	0.9963		

Dispersion parameter is fixed at 1.00.

Message: deviance ratios are based on dispersion parameter with value 1.

Message: the following units have large standardized residuals.

Unit	Response	Residual
15	1.00	2.62

Message: the residuals do not appear to be random; for example, fitted values in the range 0.04 to 0.22 are consistently larger than observed values and fitted values in the range 0.47 to 1.00 are consistently smaller than observed values.

Message: the following units have high leverage.

Unit	Response	Leverage
5	0.00	0.272
17	1.00	0.409
23	1.00	0.370

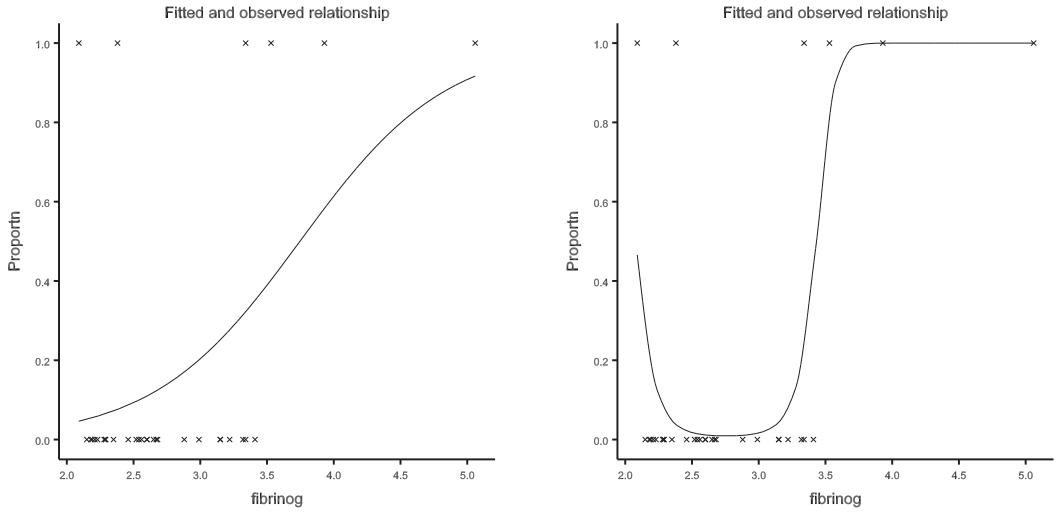
Estimates of parameters

Parameter	estimate	s.e.	t(*)	t pr.	antilog of estimate
Constant	73.4	39.2	1.87	0.061	*
fibrinog Lin	–56.6	29.3	–1.93	0.053	*
fibrinog Quad	10.28	5.27	1.95	0.051	29064.

Message: s.e.s are based on dispersion parameter with value 1.

Carry out a significance test to compare model D with model A. What do you conclude? [*3*]

(g) Figures 5(a) and 5(b) show the fitted model plots of models A and D, respectively.



(a) Model A (b) Model D

Figure 5

Analyse Figure 5, describing the fit of the fitted model curves to the data, and identify the points flagged as having large residuals. State which model you think represents the best fit to the data on the basis of these plots. [*4*]

(h) The two points flagged as having large residuals in model A were temporarily removed from the dataset. Logistic regression models that included as the explanatory variables just fibrinog (model A2) and added to this a quadratic term for the square of fibrinog (model D2), were fitted to this reduced dataset. Summary of analysis tables were obtained from GenStat as follows.

Model A2 (same as model A, but with units 15 and 23 removed)

Summary of analysis

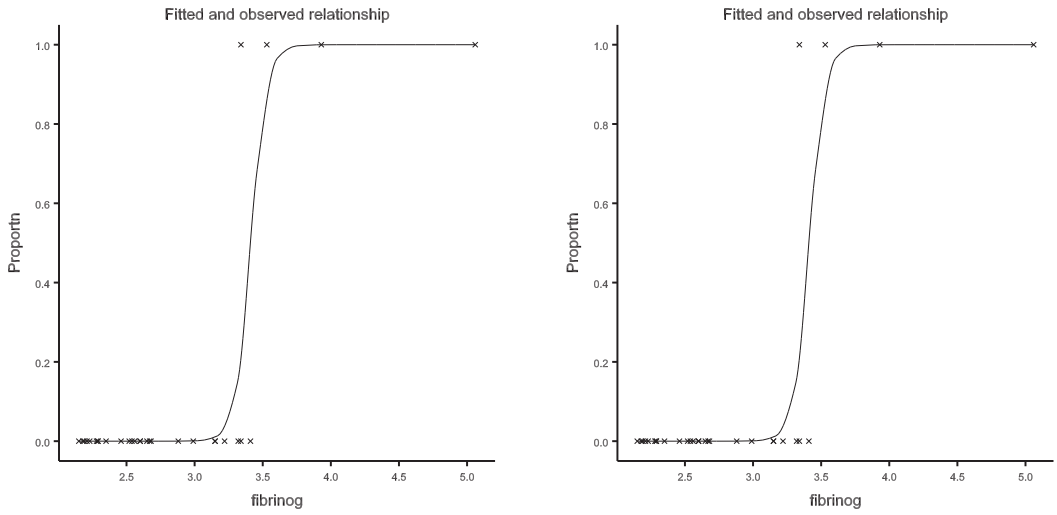
Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	1	17.978	17.9782	17.98	<.001
Residual	28	5.582	0.1994		
Total	29	23.560	0.8124		

Model D2 (same as model D, but with units 15 and 23 removed)

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	2	17.981	8.9905	8.99	<.001
Residual	27	5.579	0.2066		
Total	29	23.560	0.8124		

Figures 6(a) and 6(b) show fitted model plots for models A2 and D2, respectively.



(a) Model A2

(b) Model D2

Figure 6

Compare this output with that from models A and D. Comment on Figure 6. Perform another significance test comparing models A2 and D2, fitted to this reduced dataset. [5]

(i) As a result of what you know about the background to these data and the data analyses shown, would you say that it is sensible to include a quadratic term in the model (for the full data)? Briefly explain your answer. [3]

You should be able to answer this question after you have studied Unit 9.

Question 3 – 33 marks

(a) Waves can cause damage to the forward section of certain cargo-carrying vessels, and data on such damage were recorded. The chance that a ship will incur damage (and it could incur it more than once) is influenced by the length of time that the ship has been in service. It may also be influenced by what type of ship it is, its year of construction, and when it was in operation. Data were grouped together so that ships in each group were of the same type and were constructed in the same period. The number of damage incidents incurred by each group was recorded for each of two time periods, together with their aggregate number of months in service in those periods.

These data are given in the file `ShipDamage.gsh`, which has 34 rows, one for each observed combinations of type of ship, year of construction and period of operation. Information on five variables is given, coded as follows.

ship	Ship type, coded 1 to 5.
construct	Year of construction, coded 1: 1960–4, 2: 1965–9, 3: 1970–4, 4: 1975–9.
operation	Period of operation, coded 1: 1960–74 and 2: 1975–9.
service	Aggregate number of months of service, ranging from 45 to 44 882.
incidents	Number of damage incidents, ranging from 0 to 58.

[Source: McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, 2nd edition, Chapman and Hall, p. 204.]

A generalised linear model is needed for predicting the number of damage incidents from the other variables.

(i) From the nature of the data and by drawing a histogram for the response variable, explain why a Poisson distribution seems a reasonable distribution to assume for the response. [*3*]

(ii) Fit a generalised linear model that has incidents as the response and ship, construct, operation and service as explanatory variables. What is the residual mean deviance for the fitted model? Do any of the *Message*s from GenStat suggest problems with the model? [*5*]

(iii) Construct a new variable called lservice that equals the (natural) log of service. Give the first three values of lservice. [*1*]

(iv) Fit the same generalised linear model that you fitted in part (a)(ii), but with service replaced by lservice as one of the explanatory variables. Explain whether this is a better model than the one that you fitted in part (a)(ii). (Note: In *Unit 12* you will learn that it is better to treat lservice as an *offset* rather than an ordinary explanatory variable. However, ignore that here.) [*6*]

(v) Use the Fitted Model button in the Regression Further Output dialogue box to produce a scatterplot with lservice as the Explanatory Variable and construct as the Grouping Factor. Comment on the fit of the model. [*2*]

(vi) Try dropping each of ship, construct and operation in turn from the model fitted in part (a)(iv) (while leaving the other explanatory variables in the model). For each case give the deviance, mean deviance and approx chi pr for the change in model, and decide whether the explanatory variable can be left out of the model. Which is the best model? [*8*]

(b) Doses of insulin were given to mice, and whether or not a mouse had a positive response was recorded. Nine different dose levels were used, and each level was given to between 30 and 38 mice. Table 7 gives the logarithm of the dose level (ldose), the number of mice that received that dose level (nmice), and the number of those mice that had a positive response (nresponse). A binomial regression model is required that relates the probability of a positive response to ldose.

Table 7

ldose	nmice	nresponse
0.53	33	0
0.72	32	5
0.85	38	11
0.93	38	14
1.02	34	16
1.11	37	21
1.26	31	23
1.32	37	30
1.45	30	27

[Source: Finney, D.J. (1964) *Statistical Method in Biological Assay*, London, Griffin.]

(i) Type the data into a GenStat spreadsheet in a form suitable for fitting a binomial regression model. Print out the spreadsheet and send it to your tutor. [*2*]

(ii) Fit the binomial regression model. What are estimates of its parameters? Describe the relationship between dose and response. [*4*]

(iii) What is the outcome of the test that there is no regression effect? [*2*]

Questions 1 to 5 below form Tutor-marked Assignment M346 04. Do Questions 1 to 3 and *either* Question 4 *or* Question 5. Questions 1 to 3 are on *Units 10* to *12*, while Questions 4 and 5 are essay questions of a more general nature. Each question is marked out of 25. (The whole TMA is marked out of 100.)

The new datafiles used in this TMA and instructions for their installation (if you have not already installed them) are available from the M346 website.

You should be able to answer this question after you have studied Unit 10.

Question 1 – 25 marks

The data under investigation in this question are to be found in the file `solea.gsh`. They come from a study of the environmental factors determining the distribution of sole (*Solea solea*), a flatfish, in the Tagus estuary in Portugal. A total of 65 samples were taken from different areas in the estuary in 1995 and 1996, using a trawl. In each case, a record was made of whether this particular flatfish was present. This is recorded in the variate presabs in the datafile, coded such that 1 means that the fish was present and 0 means that it was absent. The other variables in the file are the water temperature in °C (temp), the salinity of the water in parts per thousand (sal), the percentage of gravel in the sediment (gravel), and a factor giving the month when the sample was taken (month). The overall aim was to investigate the relationship between these variables and the presence or absence of the flatfish. (Other variables were also recorded in the original study but were removed from consideration after preliminary modelling and are not included in the datafile.)

[Source: Zuur, F.A., Ieno, E.N. and Smith, G.M. (2007) *Analysing Ecological Data*, Springer, Chapter 21.]

(a) Produce boxplots and scatterplots to describe the relationship between the response variate presabs and each of the first three explanatory variables (temp, sal and gravel) in turn. (To produce these plots, you may have to temporarily change variates to factors, or vice versa.) Decide whether boxplots or scatterplots are more informative about the relationships, and send those plots (and not the less informative type of plot) to your tutor. Comment on what the plots show. [*6*]

(b) Since the response is binary, an appropriate model for the data would seem to be a generalised linear model with a Bernoulli (i.e. binomial with $n = 1$) response distribution. Fit such a model (using its canonical link function) to the data in GenStat, using all four explanatory variables (but no interactions). Include the resulting output in your answer. Produce a composite residual plot, and comment on whether the plots cast doubt on the model assumptions. Also, explain why the apparent strong pattern in the plot of residuals against fitted values is not in itself a cause for concern. Produce index plots of leverage values and Cook statistics. Which observations (if any) have high leverage? There are three observations that have relatively high Cook statistics and are thus influential. For those observations that have high influence, explain whether this is due to high leverage, large residuals, or a combination of these two causes. [*11*]

(c) Remove from consideration the three most influential points that you found in part (b). (It is not easy to see from the index plot exactly which these are; they are the points in rows 34, 58 and 65.) Re-fit the same binary regression model to the remaining data. Compare the fitted models that you obtained here and in part (b), and say (without using any formal test) whether you think that the three omitted points were (taken together) making an important difference to the model. Perform the same range of diagnostics that you did in part (b), but this time report to your tutor only the most important features of your analysis. [*8*]

You should be able to answer this question after you have studied Unit 11.

Question 2 – 25 marks

Table 8 contains a multiway contingency table describing the characteristics of 4831 car accidents. The data are also provided in the file `CarAccidents.gsh`.

The accidents were classified according to type of accident, severity of accident, type of car (S = small, C = compact, St = standard), and whether or not the driver was ejected. You are asked to use log-linear models to investigate the patterns of association in this table.

Table 8

Accident type	Accident severity	Driver ejected	Type of car		
			S	C	St
Collision with vehicle	Not severe	No	95	166	1279
		Yes	8	7	65
	Moderately severe	No	31	34	506
		Yes	2	5	51
	Severe	No	11	17	186
		Yes	4	5	54
Collision with object	Not severe	No	34	55	599
		Yes	5	6	46
	Moderately severe	No	8	34	241
		Yes	2	4	26
	Severe	No	5	10	89
		Yes	0	1	30
Rollover without collision	Not severe	No	23	18	65
		Yes	6	5	11
	Moderately severe	No	22	17	118
		Yes	18	9	68
	Severe	No	5	2	23
		Yes	5	6	33
Other rollover	Not severe	No	9	10	83
		Yes	6	2	11
	Moderately severe	No	23	26	177
		Yes	13	16	78
	Severe	No	8	9	86
		Yes	7	6	86

[Source: Fienberg, S.E. (1981) *The Analysis of Cross-classified Categorical Data*, Cambridge, MA, MIT Press, p. 91.]

(a) Load the data into GenStat and create a contingency table, classified by the factors accident, severity, car and ejected, that is similar to Table 8. Print out your table to send to your tutor. [2]

(b) Fit a log-linear model to the data that contains the four main effects accident, severity, car and ejected, but no interactions. Include the output in your answer. [2]

(c) If the main effects model from part (b) fitted the data adequately, what would that tell you about the relationship between accident, severity, car and ejected? Say whether you think the model fits the data adequately, giving two reasons for your answer (no formal statistical test is required). [4]

(d) A log-linear model is to be fitted to these data that contains accident, severity, car and ejected as main effects and all their two-factor interactions. Fit this model but do not include the output in your answer. Conduct a formal statistical test to decide whether the model fits the data adequately. You should state the value of the residual deviance, say what its distribution is under the null hypothesis, state the resulting p value, and give your conclusion. [4]

(e) Try dropping each two-way interaction from the model including all two-way interactions (while leaving the other interactions in the model). In each case decide whether the interaction can be left out of the model. State the p value for each test (or say if the p value is less than 0.001), and summarise your conclusions from the test. Which is the best model? [8]

(f) By making use of the best model found in part (e), calculate the probability that a driver will be ejected from a standard car in a severe rollover without collision, i.e. the conditional probability

$$P(\mathsf{ejected} = 1 \mid \mathsf{accident} = 3,\ \mathsf{car} = 3,\ \mathsf{severity} = 3).$$ [5]

You should be able to answer this question after you have studied Unit 12.

Question 3 – 25 marks

In 1993, a survey of bicycle and other traffic was carried out in the area round the campus of the University of California at Berkeley. City blocks in the area were classified into six groups on the basis of two factors: the type of block (busy, fairly busy or residential streets) (type) and whether or not there were bike routes (bikert). A random sample of ten blocks from each of these groups was taken. Each block was observed for one hour, and the numbers of bicycles (bikes) and of other vehicles (other) travelling along the block were recorded. (Data for two of the residential blocks without bike routes were subsequently lost.) The total number of vehicles (total) passing each block during the survey period was calculated by summing the number of bikes and other vehicles.

[Source: Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian Data Analysis*, 2nd edition, Boca Raton, Chapman & Hall/CRC, p. 98.]

To begin with, parts (a) to (c) investigate how the proportion of bicycles in the total traffic on a block depends on the group into which the block falls.

(a) Why should a binomial regression model come immediately to mind for these data? [*1*]

(b) A binomial regression model was fitted in GenStat using the canonical (logit) link, with type*bikert in the Model to be Fitted field. The following GenStat output was obtained (model A).

Model A

Regression analysis

Response variate:	bikes
Binomial totals:	total
Distribution:	Binomial
Link function:	Logit
Fitted terms:	Constant + bikert + type + bikert.type

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	5	1043.8	208.768	208.77	<.001
Residual	52	498.8	9.592		
Total	57	1542.6	27.064		

Dispersion parameter is fixed at 1.00.

Message: deviance ratios are based on dispersion parameter with value 1.

Message: the following units have large standardized residuals.

Unit	Response	Residual
2	9.00	–2.69
9	35.00	–3.88
10	55.00	7.35
15	9.00	–2.41
17	9.00	3.19
19	8.00	2.43
22	19.00	5.06
23	38.00	4.27
27	29.00	7.13
28	18.00	–4.53
31	5.00	–3.31
33	58.00	3.26
35	0.00	–2.41
39	60.00	–2.66
40	51.00	–3.57
41	58.00	–3.23
42	59.00	5.34
43	53.00	5.61
46	60.00	–3.60
47	71.00	7.34
48	63.00	3.40
49	8.00	–2.41
51	6.00	–4.06
52	9.00	–3.57
54	61.00	4.59
56	75.00	5.10
57	14.00	–2.61

Message: the following units have high leverage.

Unit	Response	Leverage
9	35.00	0.27
15	9.00	0.31

Estimates of parameters

Parameter	estimate	s.e.	t(*)	t pr.	antilog of estimate
Constant	–2.521	0.144	–17.49	<.001	0.08037
bikert yes	1.023	0.163	6.28	<.001	2.782
type fairbusy	–0.911	0.157	–5.81	<.001	0.4023
type busy	–1.799	0.157	–11.44	<.001	0.1655
bikert yes .type fairbusy	0.043	0.184	0.24	0.814	1.044
bikert yes .type busy	0.358	0.179	2.00	0.046	1.431

Message: s.e.s are based on dispersion parameter with value 1.

Parameters for factors are differences compared with the reference level:

Factor	Reference level
bikert	no
type	resid

What *two* features of the resulting output give evidence that there is overdispersion in these data with model A? [*2*]

(c) Models B and C fitted binomial regressions, as for model A, but making an adjustment that allows for the overdispersion in the data. Model C does not include the interaction term. GenStat gave the following output for models B and C.

Model B

Regression analysis

Response variate:	bikes
Binomial totals:	total
Distribution:	Binomial
Link function:	Logit
Fitted terms:	Constant + bikert + type + bikert.type

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx F pr.
Regression	5	1043.8	208.768	21.76	<.001
Residual	52	498.8	9.592		
Total	57	1542.6	27.064		

Dispersion parameter is estimated to be 9.59 from the residual deviance.

Message: the following units have large standardized residuals.

Unit	Response	Residual
10	55.00	2.37
47	71.00	2.37

Message: the following units have high leverage.

Unit	Response	Leverage
9	35.00	0.27
15	9.00	0.31

Estimates of parameters

Parameter	estimate	s.e.	t(52)	t pr.	antilog of estimate
Constant	–2.521	0.446	–5.65	<.001	0.08037
bikert yes	1.023	0.505	2.03	0.048	2.782
type fairbusy	–0.911	0.485	–1.88	0.066	0.4023
type busy	–1.799	0.487	–3.69	<.001	0.1655
bikert yes .type fairbusy	0.043	0.569	0.08	0.940	1.044
bikert yes .type busy	0.358	0.556	0.64	0.522	1.431

Message: s.e.s are based on the residual deviance.

Parameters for factors are differences compared with the reference level:

Factor	Reference level
bikert	no
type	resid

Model C

Regression analysis

Response variate:	bikes
Binomial totals:	total
Distribution:	Binomial
Link function:	Logit
Fitted terms:	Constant + bikert + type

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx F pr.
Regression	3	1034.5	344.830	36.65	<.001
Residual	54	508.1	9.410		
Total	57	1542.6	27.064		
Change	2	9.4	4.677	0.49	0.617

Dispersion parameter is estimated to be 9.41 from the residual deviance.

Message: the following units have large standardized residuals.

Unit	Response	Residual
47	71.00	2.48

Message: the following units have high leverage.

Unit	Response	Leverage
9	35.00	0.220

Estimates of parameters

Parameter	estimate	s.e.	t(54)	t pr.	antilog of estimate
Constant	–2.681	0.245	–10.93	<.001	0.06852
bikert yes	1.224	0.162	7.53	<.001	3.400
type fairbusy	–0.835	0.248	–3.36	0.001	0.4338
type busy	–1.532	0.234	–6.55	<.001	0.2161

Message: s.e.s are based on the residual deviance.

Parameters for factors are differences compared with the reference level:

Factor	Reference level
bikert	no
type	resid

(i) Using the information given in the Estimates of parameters table for model B, comment on whether each of the two factors type and bikert appear to have an effect on the proportion of bicycles in the traffic on a block, and also on whether these two factors interact. [3]

(ii) Why is it not possible to compare deviance differences with a χ^2-distribution when comparing the fit of models B and C? [1]

(iii) Explain how GenStat was used to fit model C after model B had been fitted. [2]

(iv) On the basis of the output from model C, what can you say about the need to include the interaction term? Could the model be simplified further? [3]

(v) Use the output from model C to calculate a point estimate for the proportion of bicycles in the traffic along a residential block that has a bike route. [3]

(vi) Two options offered by GenStat for model checking are a plot of residuals against fitted values and a half-normal plot of residuals. Of these two plots, which is likely to be more useful for checking models B and C? Explain your answer. [3]

For the rest of the question (parts (d) to (f)), we now investigate how the total traffic (bicycles plus other vehicles) on a block is related to the type of block and presence or absence of bike routes.

(d) One approach would be to analyse the data by using total as the response variate, and fitting an appropriate non-normal generalised linear model. State what response distribution you would use, giving reasons. [2]

(e) Instead of fitting a non-normal generalised linear model, it was decided to analyse the data using normal linear regression. Table 9 shows the means and variances of the total vehicle count for each of the six groups of city blocks.

Table 9

bikert	type	Mean	Variance
No	resid	87.4	5 389
No	fairbusy	878.1	197 281
No	busy	1958.3	415 332
Yes	resid	116.0	5 142
Yes	fairbusy	364.5	62 685
Yes	busy	1214.9	294 933

On the basis of this table, explain why it is necessary to transform the counts before using normal linear regression to analyse them. What transformations on the ladder of powers would it be appropriate to consider for this data? [2]

(f) It was decided that a square root transformation worked best for these data, and a new variate was created for the square root of the total counts sqrttot. GenStat's linear regression commands were used to analyse the transformed data, using type*bikert as the Model to be Fitted. The model output is given below (model D).

Model D

Regression analysis

Response variate: sqrttot
Fitted terms: Constant + bikert + type + bikert.type

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	5	9208.	1841.57	37.96	<.001
Residual	52	2523.	48.52		
Total	57	11731.	205.80		

Percentage variance accounted for 76.4
Standard error of observations is estimated to be 6.97.

Message: the following units have large standardized residuals.

Unit	Response	Residual
35	9.49	–2.87

Message: the error variance does not appear to be constant; large responses are more variable than small responses.

Estimates of parameters

Parameter	estimate	s.e.	t(52)	t pr.
Constant	8.59	2.46	3.49	0.001
bikert yes	1.85	3.30	0.56	0.578
type fairbusy	19.85	3.30	6.01	<.001
type busy	35.14	3.30	10.63	<.001
bikert yes .type fairbusy	–12.64	4.54	–2.78	0.007
bikert yes .type busy	–11.67	4.54	–2.57	0.013

Parameters for factors are differences compared with the reference level:

Factor	Reference level
bikert	no
type	resid

Would you consider leaving out either of the explanatory variables, or their interactions, from model D? Explain your answer. [*3*]

Questions 4 and 5

You should answer only one of Questions 4 and 5, both of which are essay questions.

For either question you are asked to write a short essay on topics from the module. By the word 'essay', we do not mean to imply that your answer should be entirely text; formulae and mathematical symbols, if appropriate, are allowed. However, you should think of this as an essay question in the sense of structure and readability. This question is included partly to give you practice for the examination, in which you will have to answer a question of this general type. See the Specimen Examination Papers for more details. The questions are from past examinations and do not focus on the units that you have studied recently, so they are partly revision.

Where appropriate, you should illustrate the points that you make with examples (taken from the module or elsewhere). If you use an example from the module, you need not repeat details of the analysis if it is given in the module text, but you should give a clear reference to where the relevant details are given in the text, and you should make it clear which particular aspect of the example is relevant to the point that you are making. (For the essay in the examination, you will not be expected to remember details of references like this, but they are appropriate for a TMA.)

Your essay should be no more than 650 words in length. If you express yourself clearly and concisely, you should be able to write a good answer which is shorter than that. Four marks are awarded for structure and clarity; these are awarded for putting the essay together in reasonably clear manner. The structure should include beginning, middle and conclusion, language should be clear and concise, and references should be included where necessary.

Question 4 – 25 marks

Two particular cases of generalised linear modelling are binomial regression and normal linear regression. Write a short essay in which you make clear the similarities and differences between these two methodologies.

Your essay should include:

- a brief definition of the generalised linear model (omitting any discussion of methods of inference at this stage); [*2*]
- a brief explanation of how each of binomial and normal linear regression are special cases of the generalised linear model, and the type of data for which they are suitable, stressing features specific to each model; [*8*]
- a brief description of methods of inference in the generalised linear model, including the basis of inference and how hypotheses are tested; [*6*]
- a brief explanation of how methods of inference for each of binomial and normal linear regression, in turn, relate to those for the generalised linear model. [*5*]

The remaining four marks are for the clarity and structure of your essay. [*4*]

Question 5 – 25 marks

Write a short essay about the similarities and differences between the following experimental designs:

- a completely randomised experiment with one treatment factor;
- a two-way factorial experiment;
- a randomised block design experiment with one blocking factor and one treatment factor.

In your essay, you should include the following.

- A verbal description of a completely randomised experiment with one treatment factor, and a brief explanation of why such a design might be used in preference to an observational study. [*2*]
- Verbal descriptions of a two-way factorial experiment and a randomised block design with one blocking factor and one treatment factor, and for each of these two designs, brief explanations of why such designs might be used in preference to a completely randomised design with one treatment factor. [*6*]
- An appropriate model for a completely randomised experiment with one treatment factor, with particular reference to the interpretation of the terms in the model. [*3*]
- An appropriate model for a two-way factorial experiment, and a description of how the terms in the model are interpreted. [*4*]
- A description of similarities between an appropriate model for a two-way factorial experiment and the model for a completely randomised experiment with one treatment factor. [*3*]
- An explanation of why an appropriate model for a randomised block experiment with one blocking factor and one treatment factor can be thought of as a special case of the model for a two-way factorial experiment. [*3*]

The remaining four marks are for the clarity and structure of your essay. [*4*]
