M346/B

Third Level Course Examination 2000

Linear Statistical Modelling

Monday 9 October 2000 **2.30 pm – 5.30 pm**

Time allowed: 3 hours

This examination is in **TWO** parts. Part I carries 25% of the total available marks and Part II carries 75%.

You should attempt **ONE** question from Part I: this question carries 25 marks. You should attempt **THREE** questions from Part II: the questions in this part carry 25 marks each also.

Since all questions carry the same mark, it is not unreasonable to allot the same time to each of them. However, since good answers to the questions in Part II can be attained quite quickly, do not be alarmed if you require a little extra time on the question in Part I.

This question paper is rather long because of the inclusion of tranches of GENSTAT output. **Do not let its length put you off.** In your initial reading of the paper, you will be able to either ignore or pass over very quickly all such output.

Please start each question on a new page, and cross out rough working.

At the end of the examination

Check that you have written your personal identifier and examination number on **each** answer book used. (You may well have used only one answer book.) **Failure to do so will mean that your work cannot be identified.** Place your signed desk record on top of your answer book(s) and fix them all together with the paper fastener provided.

PART I (Questions 1 and 2)

You should attempt **ONE** *question from this part of the examination, which carries 25% of the total available marks. Each question carries 25 marks. A guide to mark allocation is shown beside each question thus: [4].*

In each question in Part I you are asked to write a short essay on a topic from the course. 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 senses of structure and readability. Indeed, 4 of the 25 marks will be awarded for putting the essay together in a reasonably clear manner, including a reasonable structure with beginning, middle and conclusion, and reasonably concise use of language. References to specific data-based examples in the course are **not** *expected. However, it may be useful to illustrate points by giving special cases, perhaps in mathematical form (e.g.* $Y \sim N(0, \sigma^2)$ *is a special case of a distributional assumption, and* $\alpha + \beta_1 x_1 + \beta_2 x_2$ *is a special case of a formula for a regression mean).*

Question 1

Virtually every model considered in *M346* is a generalized linear model. Write an essay illustrating this statement by showing how the following five models fit into the generalized linear modelling framework: (I) models for binary regression, (II) Poisson regression, (III) loglinear models, (IV) multiple linear regression and (V) analysis of variance models.

Your answer should include:

- a description of the general form of the generalized linear model; [3]
- a brief description of how each of the five more specific models (I) to (V) above can, *in turn*, be considered as special cases of the generalized linear model, and what are the special circumstances in which each is appropriate (the marks are distributed evenly across all five models); [15]
- a brief explanation of how the normal distribution makes the analysis of variance easier and more exact in two of these specific models. [3]

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

Question 2

Write an essay describing the role of treatment factors in the design and analysis of experiments.

Your answer should include:

- a description of what a factor is in this context, and what a factorial experiment is; [4]
- a description of the model that is usually used for data from a two-way factorial experiment without blocking; [3]
- a description of what *main effects* and *interactions* are in this context, explaining how they relate to features of the model for a two-way factorial experiment; [4]
- a description of the complications that can arise in the analysis of data from a factorial experiment when the treatment combinations are unequally replicated; [4]
- a discussion of how blocking is used with factorial experiments. [6]

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

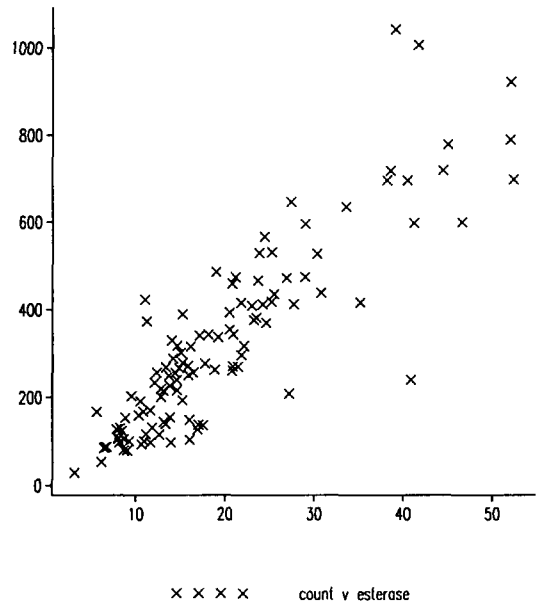
Part II (Questions 3 to 7)

You should attempt **THREE** *questions from this part of the examination, which carries 75% of the total available marks. Each question carries 25 marks. The mark allocation for each part of each question is shown beside each part thus: [4].*

Question 3

A biochemical experiment was carried out, in which the concentration of a substance called esterase was measured accurately in each of 113 samples. Then an immunological procedure was carried out on each of the samples. In this procedure a count was made of the number of 'bindings' for each of the samples. The aim of the study was to investigate the relationship between esterase concentration and the number of bindings. The data are recorded in a GENSTAT file containing variables esterase, the esterase concentration in standard units, and count, the number of bindings. Initially, the data were analysed using a normal linear regression model, with count as the response variable.

(a) The following is a scatterplot of these data.



On examining this plot, the statisticians involved decided to transform the data before further analysis. Briefly explain which feature or features of the plot would have led them to this decision. On the basis of the plot, do you think it would be appropriate to transform just one of the variables, or would it be better to transform both? Briefly explain your answer. [3]

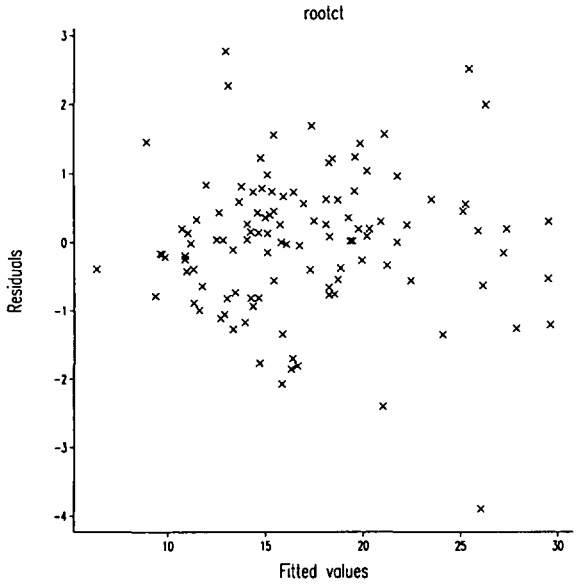
(b) The following is the output from fitting a simple linear regression model to these data in GENSTAT, after calculating a square root transformation of both variables (to give rootct as the response variable and rootest as the explanatory variable).

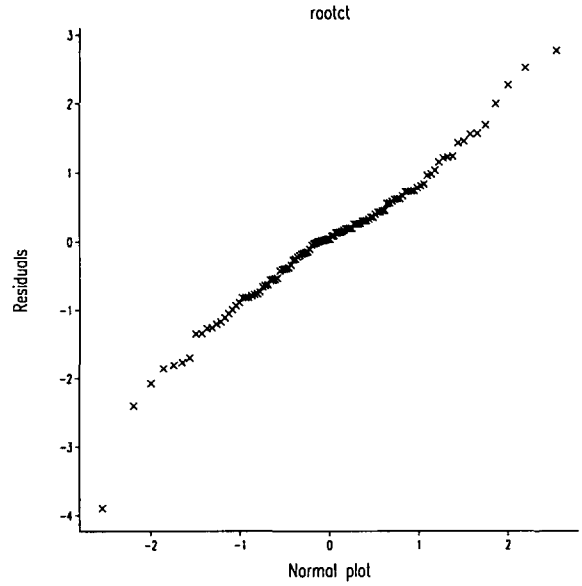
```
***** Regression Analysis *****

 Response variate: rootct
     Fitted terms: Constant, rootest


*** Summary of analysis ***

             d.f.         s.s.         m.s.      v.r.  F pr.
Regression      1       2828.0     2827.968    369.28  <.001
Residual      111        850.1        7.658
Total         112       3678.0       32.839

Change         -1      -2828.0     2827.968    369.28  <.001

Percentage variance accounted for 76.7
Standard error of observations is estimated to be 2.77
* MESSAGE: The following units have large standardized residuals:
                104       -3.90
                113        2.77
* MESSAGE: The following units have high leverage:
                 74       0.051
                 75       0.064
                106       0.047
                107       0.064
                108       0.065
                110       0.049


*** Estimates of regression coefficients ***

                estimate         s.e.    t(111)  t pr.
Constant          -1.140        0.980     -1.16  0.247
rootest            4.250        0.221     19.22  <.001
```

(i) What is the estimated regression equation resulting from this analysis? On the basis of this fitted model, calculate a point estimate for the count of bindings for a sample with an esterase concentration of 30 units. [3]

(ii) The output gives the required information to test the hypothesis that the intercept of the regression line is zero. Report what this information is, together with the results of this test, giving your conclusions clearly. [3]

(iii) The following are a plot of standardized residuals against fitted values, and a normal probability plot of residuals, for the model fitted above. Is there any feature of these plots that indicates that the assumptions of the simple linear regression model do not hold for these data? [4]





(c) An alternative approach to analysing these data is to fit a *non-normal* generalized linear model to the untransformed data.

(i) Briefly explain why Poisson regression may be a reasonable choice of model for data such as these. [2]

(ii) On the basis of the scatterplot in part (a), explain why it is likely to be inappropriate to use the canonical link function when analysing these data using Poisson regression. Suggest a link function that might be more appropriate, briefly giving a reason for your answer. [4]

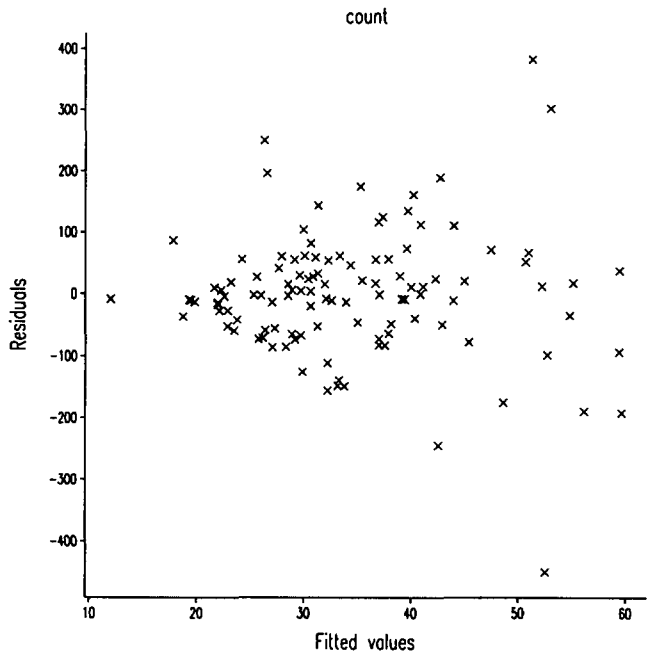
(iii) The following is an extract from a Poisson regression analysis of these data using an appropriate link function. (Details of the link function have been removed from the output.)

```
***** Regression Analysis *****

Response variate: count
    Distribution: Poisson
   Link function: XXX
    Fitted terms: Constant, esterase


*** Summary of analysis ***

                                        mean  deviance
             d.f.     deviance     deviance     ratio
Regression      1        11386.   11385.74  11385.74
Residual      111         3321.      29.92
Total         112        14706.     131.31

Change         -1       -11386.   11385.74  11385.74
* MESSAGE: ratios are based on dispersion parameter with value 1


* MESSAGE: The following units have large standardized residuals:

  <Details of 63 units with large standardized residuals appear
   here on the original printout>
```

```
* MESSAGE: The following units have high leverage:
                  75        0.048
                 107        0.047
                 108        0.048
                 110        0.148
                 112        0.050


*** Estimates of regression coefficients ***

                estimate         s.e.      t(*)
Constant         -17.44         2.74     -6.37
esterase         17.292        0.167    103.84
* MESSAGE: s.e.s are based on dispersion parameter with value 1
```

What feature (or features) of the output indicate(s) that overdispersion may be a problem? Without going into the GENSTAT details, explain briefly how you would re-analyze the data to avoid this problem. [3]

(iv) The following is a plot of unstandardized (or 'natural' in GENSTAT terminology) residuals from the above Poisson regression, against fitted values. Does the plot indicate any problems with the assumptions of the Poisson regression model? Briefly explain why or why not. [3]



Question 4

A study was made of data measured on $n = 39$ segments of large highways in the U.S. state of Minnesota in 1973. The response variable y, RATE, was the 1973 accident rate per million vehicle miles. It was of interest to relate this accident rate to a number of other variables on which data were collected. These were:

x_1, LEN: the length of the segment, in miles;

x_2, ADT: the estimated average daily traffic count, in thousands;

x_3, TRKS: the truck volume as a percentage of the total volume;

x_4, SLIM: the speed limit on that segment of road, in m.p.h.;

x_5, LWID: the lane width, in feet;

x_6, SHLD: the width of the outer shoulder of the roadway, in feet;

x_7, ITG: the number of freeway–type interchanges per mile in the segment;

x_8, SIGS: the number of signalized interchanges per mile in the segment;

x_9, ACPT: the number of access points per mile in the segment;

x_{10}, LANE: the total number of lanes of traffic in both directions;

x_{11}, FAI: 1 if the road was a federal aid interstate highway, 0 otherwise;

x_{12}, PA: 1 if the road was a principal arterial highway, 0 otherwise;

x_{13}, MA: 1 if the road was a major arterial highway, 0 otherwise.

(a) Five of the highway sections had FAI= 1, PA= 0, MA= 0, nineteen of the highway sections had FAI= 0, PA= 1, MA= 0, thirteen of the highway sections had FAI= 0, PA= 0, MA= 1, and the final two — which were classified as 'major collectors' (MC), a category different from FAI, PA and MA — had FAI= 0, PA= 0, MA= 0. How could this information have been coded more succinctly in a single factor? When and why can the version involving three variables be expected to give the same answers as the version involving a single factor? *[3]*

(b) The following GENSTAT output gives the correlation matrix of the 13 explanatory variables for this dataset, the results of a multiple regression analysis of the full 13–variable model, and the results of thirteen individual simple linear regressions of y on each of the explanatory variables in turn. (The correlation matrix and full multiple regression analysis are taken verbatim from GENSTAT; the individual simple regression results have been edited into a single table.)

```
*** Correlation matrix ***
          LEN    1.000
          ADT   -0.272    1.000
         TRKS    0.496   -0.097    1.000
         SLIM    0.186    0.244    0.296    1.000
         LWID   -0.311    0.128   -0.155    0.099    1.000
         SHLD   -0.105    0.457    0.006    0.689   -0.043    1.000
          ITG   -0.248    0.904   -0.067    0.241    0.103    0.375    1.000
         SIGS   -0.322    0.145   -0.450   -0.410    0.042   -0.134    0.070
         ACPT   -0.239   -0.224   -0.360   -0.682   -0.042   -0.425   -0.200
         LANE   -0.203    0.824   -0.153    0.265    0.096    0.482    0.698
          FAI   -0.030    0.759    0.143    0.465    0.044    0.400    0.808
           PA   -0.152    0.029   -0.052    0.044    0.225    0.367   -0.130
           MA    0.130   -0.465   -0.101   -0.424   -0.282   -0.623   -0.356

                  LEN      ADT     TRKS     SLIM     LWID     SHLD      ITG
```

SIGS	1.000					
ACPT	0.499	1.000				
LANE	0.250	-0.209	1.000			
FAI	-0.246	-0.343	0.592	1.000		
PA	0.296	-0.228	0.174	-0.374	1.000	
MA	-0.070	0.513	-0.513	-0.271	-0.689	1.000

	SIGS	ACPT	LANE	FAI	PA	MA

***** Regression Analysis *****

Response variate: RATE
Fitted terms: Constant, LEN, ADT, TRKS, SLIM, LWID, SHLD, ITG, SIGS, ACPT, LANE, FAI, PA, MA

*** Summary of analysis ***

	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	13	113.99	8.769	6.11	<.001
Residual	25	35.89	1.436		
Total	38	149.89	3.944		
Change	-13	-113.99	8.769	6.11	<.001

Percentage variance accounted for 63.6
Standard error of observations is estimated to be 1.20
* MESSAGE: The following units have large standardized residuals:

25	-2.40
27	2.44
34	-2.12

* MESSAGE: The residuals do not appear to be random;
for example, fitted values in the range 3.17 to 4.11
are consistently larger than observed values
and fitted values in the range 5.05 to 6.82
are consistently smaller than observed values

*** Estimates of regression coefficients ***

	estimate	s.e.	t(25)	t pr.
Constant	13.66	6.87	1.99	0.058
LEN	-0.0648	0.0334	-1.94	0.064
ADT	-0.0040	0.0339	-0.12	0.906
TRKS	-0.100	0.115	-0.87	0.391
SLIM	-0.1238	0.0817	-1.52	0.142
LWID	-0.134	0.598	-0.22	0.825
SHLD	0.014	0.162	0.09	0.931
ITG	-0.48	1.28	-0.37	0.714
SIGS	0.714	0.525	1.36	0.186
ACPT	0.0666	0.0426	1.56	0.130
LANE	0.027	0.284	0.09	0.926
FAI	0.54	1.73	0.31	0.756
PA	-1.01	1.11	-0.91	0.370
MA	-0.548	0.976	-0.56	0.579

*** Summary table of results of individual simple regressions on each explanatory variable in turn ***

	estimate	s.e.	t(37)	t pr.
LEN	-0.1214	0.0380	-3.20	0.003
ADT	-0.0030	0.0175	-0.17	0.863
TRKS	-0.432	0.119	-3.63	<.001
SLIM	-0.2312	0.0409	-5.66	<.001
LWID	-0.024	0.716	-0.03	0.973
SHLD	-0.2531	0.0992	-2.55	0.015
ITG	-0.120	0.794	-0.15	0.881
SIGS	1.770	0.426	4.16	<.001
ACPT	0.1603	0.0231	6.94	<.001
LANE	-0.048	0.240	-0.20	0.842
FAI	-1.217	0.943	-1.29	0.205
PA	-0.634	0.636	-1.00	0.326
MA	1.405	0.643	2.18	0.035

What does this output suggest about which explanatory variables should be in a good multiple regression model based on a subset of the thirteen available explanatory variables? [7]

(c) The stepwise regression method provided by GENSTAT, with *M346* default choices, was applied to the dataset. Both 'forwards' and 'backwards' applications of the stepwise method resulted in a model containing just x_1, LEN, x_4, SLIM and x_9, ACPT.

(i) What is meant by 'forwards' and 'backwards' applications of the stepwise method and why do they sometimes result in different models? [4]

(ii) Does the model obtained by GENSTAT seem reasonable given the preliminary analysis you did in part (b)? [2]

(d) GENSTAT output concerning the regression analysis of y, RATE, on x_1, LEN, x_4, SLIM and x_9, ACPT only is given next:

***** Regression Analysis *****

Response variate: RATE
Fitted terms: Constant, LEN, SLIM, ACPT

*** Summary of analysis ***

	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	3	105.04	35.013	27.33	<.001
Residual	35	44.85	1.281		
Total	38	149.89	3.944		
Change	-3	-105.04	35.013	27.33	<.001

Percentage variance accounted for 67.5
Standard error of observations is estimated to be 1.13
* MESSAGE: The following units have large standardized residuals:

26	2.28
27	2.26

* MESSAGE: The following units have high leverage:

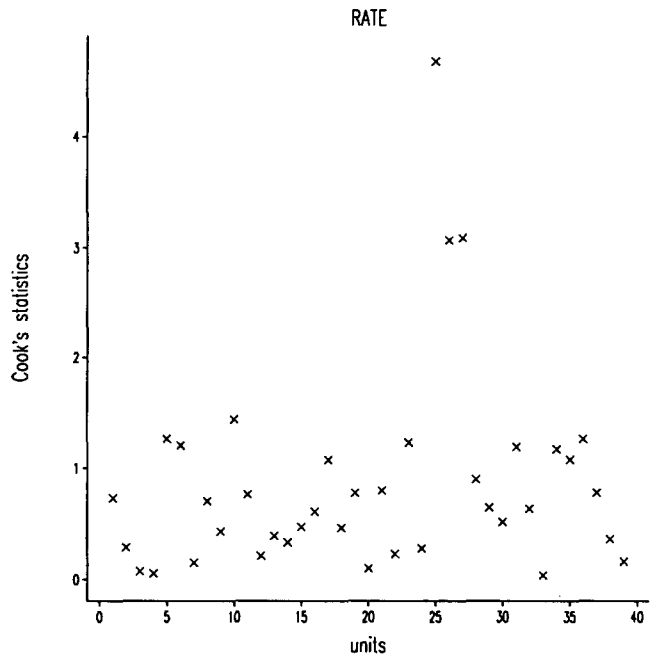
5	0.23
21	0.38
25	0.54

```
*** Estimates of regression coefficients ***

                estimate         s.e.      t(35)  t pr.
Constant            9.33         2.62       3.56  0.001
LEN              -0.0771       0.0249      -3.10  0.004
SLIM             -0.1024       0.0429      -2.39  0.023
ACPT              0.1014       0.0273       3.72  <.001
```

(i) Write down the fitted model and explain in qualitative terms what the model says about the dependence of the accident rate on explanatory variables. [3]

(ii) If a further section of Minnesota highway in 1973 was of length 12 miles, had a speed limit of 55 m.p.h., lanes of width 11 feet, and 8.5 access points per mile, what does the model predict that the accident rate per million vehicle miles would be? [2]

(e) Below is an index plot of the Cook statistics associated with the model fitted in part (d). What is the purpose of such a plot? Identify the three points that the plot draws most attention to, and describe briefly, with reference to the last tranche of GENSTAT output, what it is about these points that has made them stand out on the plot of Cook statistics.



[4]

Question 5

An experiment was conducted to investigate the effect on the growth of hybrid poplars of adding three chemicals in various combinations: lime (L), nitrogen (N) and phosphorus (P). The poplars were grown in concrete soil frames which were laid out into three blocks. Each block contained sixteen concrete frames, two for each of the eight possible treatment combinations. At the end of the experiment the weight (in grams) of the new growth in each frame after it had been dried in an oven was recorded. The data were as follows.

Chemical added	Block 1		Block 2		Block 3	
	Replicate 1	Replicate 2	Replicate 1	Replicate 2	Replicate 1	Replicate 2
none	13.9	13.6	14.3	12.7	15.8	15.5
L only	15.3	16.6	19.4	20.1	15.9	15.1
N only	57.9	31.7	21.7	25.6	31.0	25.7
P only	14.2	14.7	22.8	12.8	22.1	13.3
L and N	43.0	41.2	62.5	59.3	37.1	32.0
L and P	11.8	17.8	23.2	21.4	22.7	20.6
N and P	49.5	49.7	35.5	38.1	30.7	36.3
L, N and P	63.8	53.4	59.7	53.5	41.3	58.5

(a) (i) Ignoring the blocks, what general name can be applied to experiments with this type of treatment structure? [1]

(ii) In assigning the various treatments to frames within each block, what principles should the investigators follow, and why? [3]

(b) How would you lay out the data from the table above in a GENSTAT spreadsheet? Give the number of rows and columns, and describe the information each column would contain. [5]

(c) General analysis of variance was used in GENSTAT to fit a model (Model 1) including all the standard treatment contrasts up to and including the 3-factor interaction and a model omitting many of the interactions (Model 2). Extracts from the outputs are as follows — many numbers have been omitted from the output for Model 2.

Model 1

```
***** Analysis of variance *****

Variate: weight

Source of variation      d.f.       s.s.       m.s.      v.r.  F pr.

block stratum               2     215.45     107.72      1.83

block.*Units* stratum
L                           1     884.94     884.94     15.04  <.001
N                           1    8350.33    8350.33    141.91  <.001
P                           1     354.80     354.80      6.03  0.019
L.N                         1     395.03     395.03      6.71  0.014
L.P                         1       2.04       2.04      0.03  0.853
N.P                         1     108.30     108.30      1.84  0.183
L.N.P                       1       1.30       1.30      0.02  0.883
Residual                   38    2236.04      58.84

Total                      47   12548.22
```

Model 2
***** Analysis of variance *****

Variate: weight

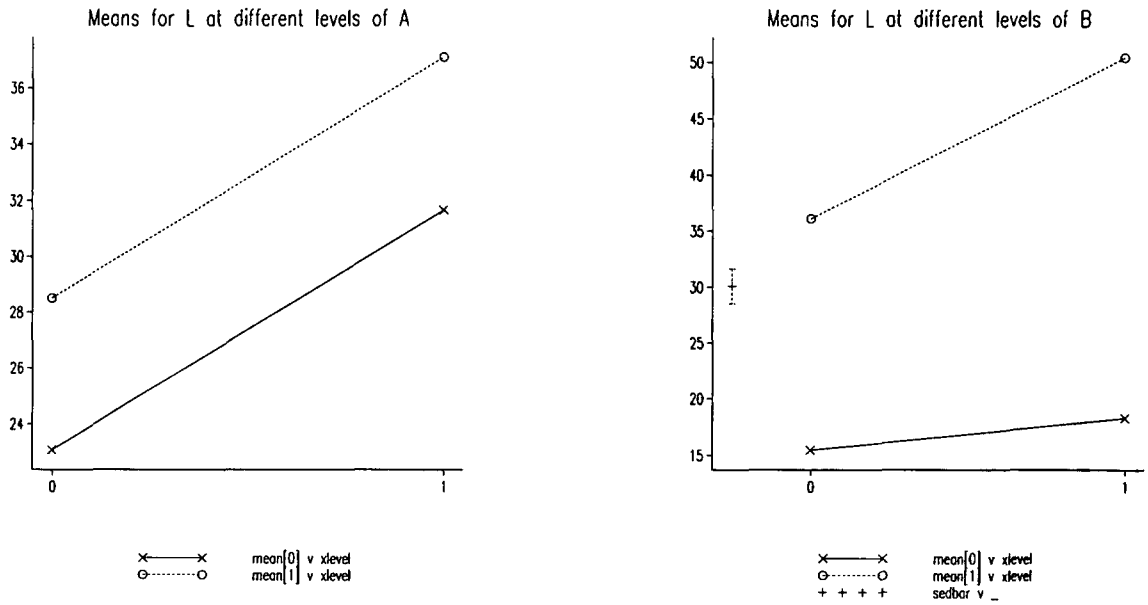
Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
block stratum					
block.*Units* stratum					
L					<.001
N					<.001
P					0.017
L.N					0.012
Residual					
Total	47	12548.22			

(i) Explain why F pr. is not given for the block stratum in the analysis of variance table for Model 1. [1]

(ii) Complete the resulting analysis of variance table for Model 2 by copying it into your answer book and filling in appropriate values in the d.f., s.s., m.s. and v.r. columns. (It may help you if you show your working.) [8]

(iii) Will the estimate of the L main effect be the same in Model 1 and Model 2? Explain why or why not. [2]

(d) After fitting Model 2, two line plots of means were drawn to illustrate the joint Lime (L) and Nitrogen (N) effects, and the joint Lime (L) and Phosphorus (P) effects. These plots appear below, but the identities of N and P have been disguised by labelling one of them as A and the other as B.



(i) Which labelling for the plots (above) appears more appropriate? Justify your answer.

Chemical A:	Nitrogen (N)	or	Chemical A:	Phosphorus (P)	[2]
Chemical B:	Phosphorus (P)		Chemical B:	Nitrogen (N)	

(ii) Assuming that for all three chemicals the following coding was used 0 – no chemical applied, 1 – chemical applied, what combination of Lime, Nitrogen and Phosphorus is the best combination to produce the most new growth? Briefly explain your answer. [3]

Question 6

Sugarcane can be propagated by tissue culture ('micropropagation') — taking small buds from existing plants, growing them first in a laboratory before transplanting the new plants outdoors. Such micropropagation has the advantage that it allows new disease-free planting material to be produced quickly. However, in micropropagation, initial contamination of the buds is a problem as it can lead to the loss of the new plants. Therefore there is interest in developing techniques to sterilise buds without damaging them in the process. As part of an experiment investigating different sterilising procedures buds were dipped in a solution of sodium hypochlorite and the number of buds that were not harmed was recorded.

Number of buds (nbud)	Concentration of solution (conc)	Length of time dipped (dip)	pH of solutions (pH)	Number of buds unharmed (nbudok)
6	2.7%	5 min	14	4
6	2.7%	10 min	14	2
7	2.7%	20 min	14	7
7	2.7%	5 min	6	0
7	2.7%	10 min	6	0
6	2.7%	20 min	6	0
6	1.7%	5 min	14	6
6	1.7%	10 min	14	6
6	1.7%	20 min	14	6
7	1.7%	5 min	6	5
6	1.7%	10 min	6	5
7	1.7%	20 min	6	4

(a) Explain why analysis of variance using percentage of buds unharmed as the dependent variable is not a good way of analysing these data. [2]

(b) Instead it was decided to fit a logistic regression model to these data using GENSTAT (Model 1). The output, together with some potentially useful χ^2 quantiles, was as follows.

Model 1

```
***** Regression Analysis *****

 Response variate: nbudok
  Binomial totals: nbud
     Distribution: Binomial
    Link function: Logit
     Fitted terms: Constant + conc + dip + pH


*** Summary of analysis ***

                                        mean  deviance
              d.f.      deviance    deviance     ratio
Regression       3           57.         19.     18.98
Residual         8            9.          1.
Total           11           66.          6.

Change          -3          -57.         19.     18.98
* MESSAGE: ratios are based on dispersion parameter with value 1


* MESSAGE: The following units have large standardized residuals:
                  2       -2.06
                  3        3.26
```

```
*** Estimates of regression coefficients ***

                              estimate         s.e.      t(*)
Constant                         0.391        0.792      0.49
conc 2.700                       -12.8         49.5     -0.26
dip                             0.0400       0.0561      0.71
pH 14                            12.7         49.5      0.26
* MESSAGE: s.e.s are based on dispersion parameter with value 1

 CUCHISQU((57; 3))
  0.2570E-11


 CUCHISQU((9; 8))
      0.3423


 CUCHISQU((66; 11))
  0.6985E-09
```

(i) In Model 1, conc and pH were treated as factors. Was dip also treated as a factor or was it treated as a variate? Justify your answer. Give one advantage and one disadvantage of treating dip this way. [4]

(ii) What general method does GENSTAT use to estimate the regression coefficients? [1]

(iii) From Model 1, what are your conclusions about the effect on bud damage of the concentration, pH and length of time dipped in sodium hypochlorite? You should mention the effect of each term individually and their effect jointly, and include SPs where possible. [3]

(c) It was decided to drop the term dip from the model to produce Model 2. The output, with some χ^2 quantiles, was as follows.

Model 2

```
***** Regression Analysis *****

 Response variate: nbudok
  Binomial totals: nbud
     Distribution: Binomial
    Link function: Logit
     Fitted terms: Constant + conc + pH


*** Summary of analysis ***

                                        mean  deviance
             d.f.      deviance    deviance     ratio
Regression      2           56.         28.     28.20
Residual        9           10.          1.
Total          11           66.          6.

Change          1            1.          1.      0.52
* MESSAGE: ratios are based on dispersion parameter with value 1


* MESSAGE: The following units have large standardized residuals:
                  2      -2.13
                  3       2.90
```

```
*** Estimates of regression coefficients ***

                              estimate         s.e.      t(*)
Constant                         0.847        0.488      1.74
conc 2.700                     -12.8          50.4      -0.26
pH 14                           12.8          50.4       0.25
* MESSAGE: s.e.s are based on dispersion parameter with value 1

 CUCHISQU((56; 2))
   0.6914E-12


 CUCHISQU((1; 1))
      0.3173
```

(i) For these data, is Model 2 better than Model 1? Justify why or why not. [3]

(ii) Using Model 2, write down the estimate of the log odds ratio (relative to a concentration of 1.7% and pH of 6) for each combination of concentration and pH that was used in the experiment. [4]

(iii) From your answer in part (c)(ii), rank the treatments in order of increasing harmfulness to the buds. [2]

(d) (i) Consider the two messages after the 'Summary of analysis' table in the output from Model 2. Does either message suggest that the model is not fitting very well? Why or why not? [4]

(ii) Does overdispersion appear to be a problem in this model? Justify your answer. [2]

Question 7

Physical anthropologists carried out a survey of three different populations of native peoples from the Arctic areas of North America. The aim was to investigate the incidence of a condition known as *torus mandibularis*, a protuberance in the lower jaw. The anthropologists chose samples of people from each of the three populations. The numbers of people sampled from each population were fixed in advance. For each person in the samples, the anthropologists recorded their sex (GENSTAT factor sex, recorded as male or female), their age (age classified into six groups, corresponding to ages (in years) 1–10, 11–20, 21–30, 31–40, 41–50, and 51 and over), and whether *torus mandibularis* was present or absent (factor presabs, the response variable). As well as sex, age and presabs, the GENSTAT file containing these data includes the factor popn (to indicate the particular population, at three levels, Iglook, Beech and Aleut). These four factors define a four-way contingency table, and the GENSTAT file includes the counts in each cell of this table in a variate count. None of these cell counts takes the value zero.

The data were analysed by using GENSTAT to fit certain log-linear models.

(a) Explain briefly why it would not be statistically legitimate to fit a model to these data that omitted the main effect of popn. [2]

(b) Appended below are the Summary of analysis tables arising from the fit of three different models to these data, together with some useful χ^2 quantiles.

```
***** Regression Analysis *****
 Response variate: count
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant + popn + sex + age + presabs + popn.sex +
                   popn.age + sex.age + popn.presabs + sex.presabs +
                   age.presabs + popn.sex.age + popn.sex.presabs +
                   popn.age.presabs + sex.age.presabs

*** Summary of analysis ***
                                          mean  deviance
              d.f.     deviance     deviance     ratio
Regression      61       488.81        8.013      8.01
Residual        10        21.61        2.161
Total           71       510.41        7.189

Change         -61      -488.81        8.013      8.01
* MESSAGE: ratios are based on dispersion parameter with value 1

***** Regression Analysis *****
 Response variate: count
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant + popn + sex + age + presabs + popn.sex +
                   popn.age + sex.age + popn.presabs + sex.presabs +
                   age.presabs

*** Summary of analysis ***
                                          mean  deviance
              d.f.     deviance     deviance     ratio
Regression      34       452.78       13.317     13.32
Residual        37        57.64        1.558
Total           71       510.41        7.189

Change         -34      -452.78       13.317     13.32
* MESSAGE: ratios are based on dispersion parameter with value 1
```

```
CUCHISQU((57.64; 37))
    0.01646

***** Regression Analysis *****

 Response variate: count
     Distribution: Poisson
     Link function: Log
     Fitted terms: Constant + popn + sex + age + presabs

*** Summary of analysis ***

                                      mean  deviance
              d.f.     deviance     deviance     ratio
Regression       9        284.7       31.631     31.63
Residual        62        225.7        3.641
Total           71        510.4        7.189

Change          -9       -284.7       31.631     31.63
* MESSAGE: ratios are based on dispersion parameter with value 1

 CUCHISQU((225.7; 62))
          0
```

(i) Summarize briefly which terms are included in the three models fitted in the GENSTAT output. [3]

(ii) Consider testing whether the `popn.sex.age.presabs` interaction can be omitted from a model that otherwise includes all main effects and interactions. What is the value of the test statistic for this test, and what distribution should it be compared with? Explain briefly why the number of degrees of freedom takes the value that it does. The SP turns out to be 0.01722. What do you conclude? [4]

(iii) Suppose that you wished to fit a non-saturated model to these data. Which is the highest level of interaction that you would choose to retain in your model? Give brief reasons for your answer. [5]

(c) A statistician carried out further analyses of these data using GENSTAT. The `Summary of analysis` table corresponding to the final model that was fitted is given below, together with a relevant χ^2 quantile and certain tables of totals.

```
***** Regression Analysis *****

 Response variate: count
     Distribution: Poisson
     Link function: Log
     Fitted terms: Constant + popn + sex + age + presabs + age.presabs

*** Summary of analysis ***

                                      mean  deviance
              d.f.     deviance     deviance     ratio
Regression      14       435.11       31.080     31.08
Residual        57        75.30        1.321
Total           71       510.41        7.189

Change         -14      -435.11       31.080     31.08
* MESSAGE: ratios are based on dispersion parameter with value 1

 CUCHISQU((75.3; 57))
     0.05265
```

popn	Total Iglook	Beech	Aleut
presabs			
absent	190.00	74.00	70.00
present	125.00	44.00	38.00

sex	Total male	female
presabs		
absent	174.0	160.0
present	116.0	91.0

age	Total 1-10	11-20	21-30	31-40	41-50
presabs					
absent	131.00	95.00	65.00	22.00	13.00
present	15.00	30.00	47.00	50.00	30.00

age	51+
presabs	
absent	8.00
present	35.00

(i) Does the model fit adequately? Briefly explain why or why not. [2]

(ii) Describe in words the model that has been fitted. What does the response variable presabs depend on, according to the fitted model? Using relevant information from the tables given, describe the main features of this dependence. [5]

(d) These data could alternatively have been analysed in GENSTAT by using logistic regression to regress the response variable on the three explanatory variables involved (population, sex and age). Briefly describe how the data would need to be entered in a GENSTAT spreadsheet to carry out such an analysis. Would you expect the results of the logistic regression and the loglinear model fitted in part (c) to be in exact correspondence? Give a reason for your answer. [4]

[END OF QUESTION PAPER]