M346/R

The Open University

Third Level Course Examination 1999
Linear Statistical Modelling

Wednesday 20 October 1999 10.00 am – 1.00 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

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

Your answer should include:

- a brief description of what a block is in this context; [2]
- an outline of the reasons for using blocks in the design of an experiment; [4]
- a brief description of at least one experimental situation where more than one kind of block is involved; [6]
- a brief explanation of how blocks are taken into account in standard models for experimental data (giving some details for at least one such model); [5]
- a brief explanation of the reason why the ANOVA commands in software like GENSTAT do not produce *SP* values for blocks. [4]

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

Question 2

In many data analyses involving linear and generalized linear models, some or all of the variables are transformed. Write an essay outlining the role of transformations in linear and generalized linear modelling.

Your answer should include:

- a definition of transformation; [2]
- an outline of reasons for transforming explanatory variables and for transforming response variables; [9]
- an explanation of how the diagnostic information and diagnostic plots produced by software like GENSTAT can indicate that a transformation is called for. [10]

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

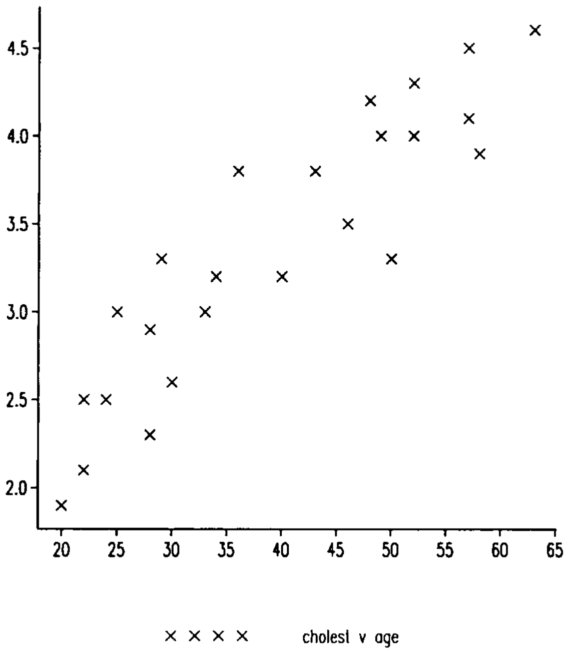
Part II (Questions 3 to 7)

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

Data were collected on 24 patients with hyperlipoproteinaemia, a condition characterised by high levels of substances called lipoproteins in the blood. Among the items recorded for each patient were the total plasma cholesterol level (in mg/ml, variable `cholest`) and the patient's age in years (variable `age`). The medical staff who collected the data were interested in modelling the relationship between these two variables, treating `cholest` as the response variable.

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



Judging from this diagram, would you say it is reasonable to fit a simple linear regression model as a first step in analysing these data? Briefly explain why or why not. [3]

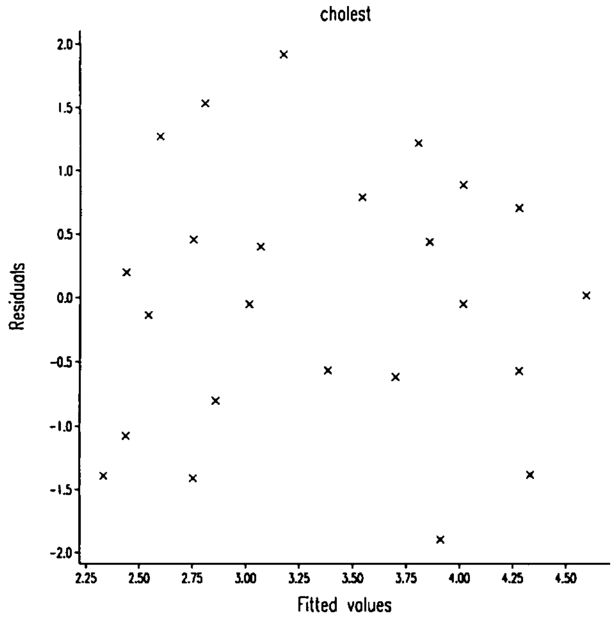
(b) The following is the output from fitting a simple linear regression model to these data in GENSTAT.

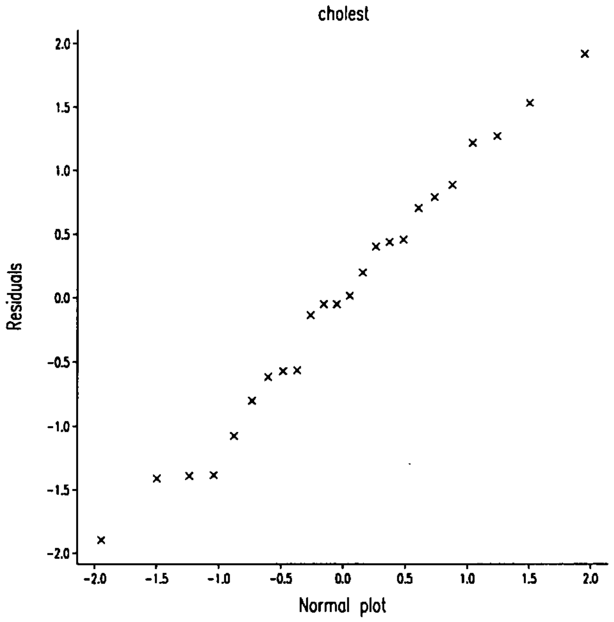
```
***** Regression Analysis *****

 Response variate: cholest
     Fitted terms: Constant, age


*** Summary of analysis ***

               d.f.         s.s.         m.s.      v.r.  F pr.
Regression       1       11.465      11.4648    102.75  <.001
Residual        22        2.455       0.1116
Total           23       13.920       0.6052

Change          -1      -11.465      11.4648    102.75  <.001

Percentage variance accounted for 81.6
Standard error of observations is estimated to be 0.334
* MESSAGE: The following units have high leverage:
                  10       0.176


*** Estimates of regression coefficients ***

                  estimate         s.e.     t(22)  t pr.
Constant             1.280        0.216      5.93  <.001
age                0.05262      0.00519     10.14  <.001
```

(i) What is the estimated regression equation resulting from this analysis? What, according to the fitted model, is the mean total plasma cholesterol level for patients with hyperlipoproteinaemia aged 60 years? [2]

(ii) GENSTAT's PREDICT command gives the mean total plasma cholesterol levels of patients aged 50 years as 3.9111 mg/ml with a standard error of 0.0876. Assuming the model is correct, calculate a 95% prediction interval for the total plasma cholesterol level of a patient aged 50 years. (The 0.975 quantile of a t distribution with 22 degrees of freedom is 2.074.) [6]

(c) The following are a plot of residuals against fitted values and a normal probability plot of residuals, for the model fitted in part (b).

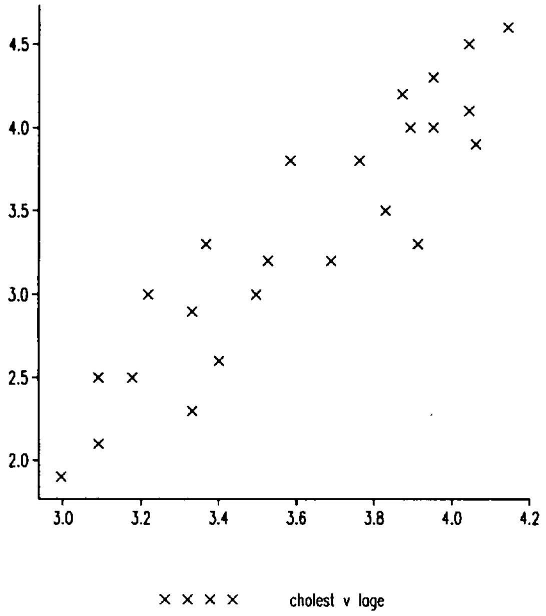




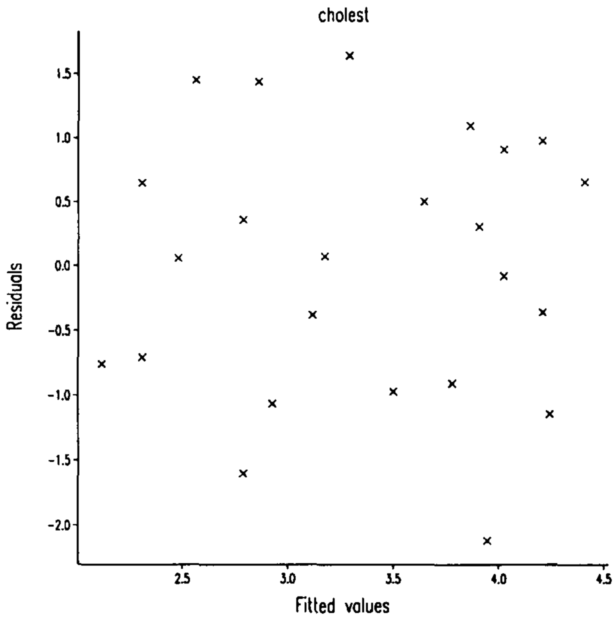
(i) Is there any feature of these plots that indicates that the assumptions of the simple linear regression model do not hold for these data? [4]

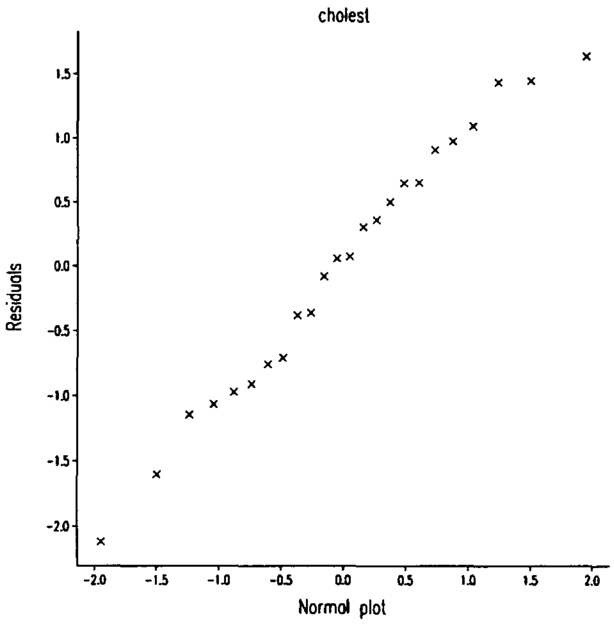
(ii) The rightmost point in the plot of residuals against fitted values corresponds to patient 10, who was flagged in the output in part (b). Briefly explain why this patient turns out nevertheless to have a low Cook statistic. [4]

(d) The statistician who analysed these data also carried out the same analysis using the logarithm (base e) of patient's age as the explanatory variable (called `lage`), but leaving the response variable untransformed. A scatter diagram, the GENSTAT output and residual plots were as follows.



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

 Response variate: cholest
     Fitted terms: Constant, lage


*** Summary of analysis ***

              d.f.        s.s.         m.s.      v.r.  F pr.
Regression       1      11.715      11.7149    116.90  <.001
Residual        22       2.205       0.1002
Total           23      13.920       0.6052

Change          -1     -11.715      11.7149    116.90  <.001

Percentage variance accounted for 83.4
Standard error of observations is estimated to be 0.317
* MESSAGE: The following units have large standardized residuals:
                  24      -2.12
* MESSAGE: The following units have high leverage:
                   7       0.172


*** Estimates of regression coefficients ***

                  estimate         s.e.     t(22)  t pr.
Constant            -3.858        0.670     -5.76  <.001
lage                 1.995        0.185     10.81  <.001
```





If the main aim of the analysis were to predict the individual total plasma cholesterol levels of new patients (on the basis of their age), would you prefer the model here or the model in part (b)? Briefly explain your answer. You should describe any disadvantages of the choice you make, as well as its advantages. [6]

Question 4

The Greensand Ridge relay is an annual cross-country relay race run along the Greensand ridge in Bedfordshire. The race distance is split in six parts (legs) each of different length ranging from 6.7km to 12.8km (Table 1).

Table 1

leg	distance (km)
1	7.5
2	8.5
3	9.4
4	12.8
5	6.7
6	8.7
Overall	**53.6**

The race is basically a "fun" event. It attracts both male and female competitors of all ages. There is therefore interest in producing a fair handicapping system so that no team is more likely to win simply because of its age/gender composition. The handicapping system works by giving teams appropriate head starts so that every team is expected to arrive at the finish at the same time. It is therefore important to be able to model how long a competitor is expected to take to run a leg.

The results from a previous year's Greensand Ridge relay (kindly collated by Roger Williams, South Midlands Orienteering Club) were entered on to a GENSTAT spreadsheet. In view of previous experience with data of this kind, a quadratic term in age was included. The columns were as follows:

`age` : age (in years) – 40
`age2` : age^2
`gender` : the gender of the competitor (0 – male, 1 – female)
`dist` : length of the leg run in km
`time` : time taken to run the leg in minutes

(a) The scatterplot matrix of all the variables is given in Figure 1. What are your conclusions on the basis of these plots? What limitations do these particular plots have? [5]

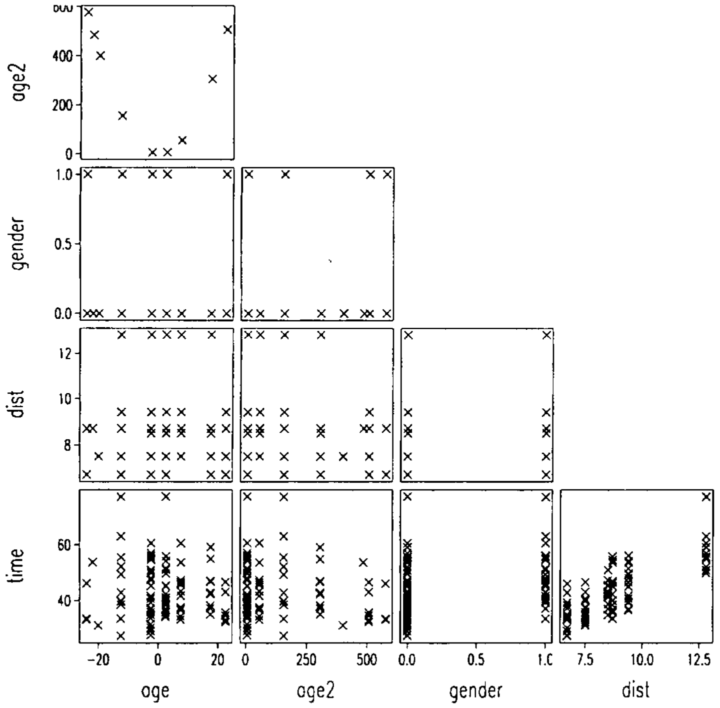


Figure 1

(b) A multiple linear regression analysis including **age**, **age2**, **dist** and **gender** was performed using GENSTAT (Model 1). From the output given below, what are your conclusions about allowing for the competitor's age in the handicapping system? [3]

Model 1

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

Response variate: time
     Fitted terms: Constant + age + age2 + dist + gender

*** Summary of analysis ***

              d.f.        s.s.        m.s.     v.r.  F pr.
Regression       4       6358.     1589.59    56.98  <.001
Residual        85       2371.       27.90
Total           89       8730.       98.09

Change          -4      -6358.     1589.59    56.98  <.001

Percentage variance accounted for 71.6
Standard error of observations is estimated to be 5.28
* MESSAGE: The following units have large standardized residuals:
                  59        2.57
                  60        3.96
                  88        2.82
* MESSAGE: The following units have high leverage:
                  14        0.165
                  66        0.190
                  75        0.191
                  76        0.183

*** Estimates of regression coefficients ***

                        estimate          s.e.     t(85)  t pr.
Constant                    4.40          2.84      1.55  0.125
age                       0.0787        0.0535      1.47  0.145
age2                     0.00324       0.00331      0.98  0.330
dist                       4.100         0.294     13.94  <.001
gender                      7.88          1.33      5.93  <.001
```

(c) Stepwise regression in GENSTAT was then used to find a more parsimonious model (Model 2 given below).

Model 2

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

 Response variate: time
     Fitted terms: Constant + dist + gender

*** Summary of analysis ***

             d.f.         s.s.         m.s.      v.r.
Regression     2         6262.      3131.21    110.41
Residual      87         2467.        28.36
Total         89         8730.        98.09

Change        -2        -6262.      3131.21    110.41

Percentage variance accounted for 71.1
Standard error of observations is estimated to be 5.33
* MESSAGE: The following units have large standardized residuals:
                  60        3.89
* MESSAGE: The following units have high leverage:
                  50        0.087
                  56        0.087
                  58        0.087
                  59        0.087

*** Estimates of regression coefficients ***

                              estimate         s.e.     t(87)
Constant                          5.92         2.68      2.21
dist                             4.010        0.290     13.82
gender                            7.12         1.27      5.61
```

(i) Briefly describe in general the process by which GENSTAT arrives at a parsimonious model using stepwise regression. [4]

(ii) For model 2, describe what a graph of the fitted values against **dist** would look like. [3]

(iii) Using model 2, how long would a team with 4 men (on legs 1 to 4) and 2 women (on legs 5 and 6) be expected to take for the whole race next year? [3]

(iv) What sources of uncertainty are there for your estimate in part (c)(iii)? [3]

(d) The current handicapping system for the Greensand ridge relay is based on modelling the reciprocal of speed (i.e. time/distance) of competitors. Explain why multiple linear regression using the reciprocal of speed as the response variable and **age**, **age2** and **gender** as explanatory variables cannot be equivalent to either Model 1 or Model 2. [4]

Question 5

A group of chemical engineers believed that the time required to complete a chemical reaction was affected by the amounts of two chemicals A and B, and by the temperature C. They wished to find new experimental settings that yielded a shorter reaction time while maintaining low costs.

Their current experimental settings were coded as A = 1, B = 1 and C = 1, and they decided to try changing the current settings by decreasing the values of the corresponding quantities because, if any of the new changes appeared to be effective, the current costs would be either maintained or decreased. They coded the new values as A = −1, B = −1 and C = −1.

To test the effect of changing the current settings on the reaction time, they designed an experiment in which two reaction times were observed at each of the 8 possible combinations of values of the amounts of chemicals A and B and the temperature C. Table 2 gives the two replicates of the reaction time observed for each of the 8 treatment combinations.

Table 2

	Factors			time	
Treatment	A	B	C	y_{i1}	y_{i2}
1	1	1	1	10	12
2	1	1	−1	10	8
3	1	−1	1	4	5
4	1	−1	−1	11	13
5	−1	1	1	10	8
6	−1	1	−1	7	9
7	−1	−1	1	5	3
8	−1	−1	−1	8	8

The data collected from this experiment were stored in a GENSTAT spreadsheet with four columns A, B, C, all factors, and time, a variate) and 16 rows.

(a) How would you describe the experimental design? [1]

(b) The experimental data were analysed using the GENSTAT analysis of variance commands, by selecting time as the Y-Variate and A*B*C as the Treatment Structure. Some of the output produced is given below (Model 1).

Model 1

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

Variate: time

Source of variation     d.f.       s.s.       m.s.     v.r.  F pr.
A                          1     14.063     14.063     9.00  0.017
B                          1     18.062     18.062    11.56  0.009
C                          1     18.062     18.062    11.56  0.009
A.B                        1      0.562      0.562     0.36  0.565
A.C                        1      1.562      1.562     1.00  0.347
B.C                        1     52.563     52.563    33.64  <.001
A.B.C                      1      5.062      5.062     3.24  0.110
Residual                   8     12.500      1.562
Total                     15    122.438
```

```
***** Tables of means *****

Variate: time

Grand mean  8.19

        A       -1        1
              7.25     9.13

        B       -1        1
              7.12     9.25

        C       -1        1
              9.25     7.12

        A        B       -1        1
       -1              6.00     8.50
        1              8.25    10.00

        A        C       -1        1
       -1              8.00     6.50
        1             10.50     7.75

        B        C       -1        1
       -1             10.00     4.25
        1              8.50    10.00

                 B       -1                 1
        A        C       -1        1       -1        1
       -1              8.00     4.00     8.00     9.00
        1             12.00     4.50     9.00    11.00
```

Use the output to answer the following questions.

(i) Without performing a formal statistical test, say whether the output provides evidence that any of the treatment combinations has an effect on the reaction time. Explain your answer briefly. [3]

(ii) Conduct a formal statistical test to check whether there is evidence that adding the three-factor interaction A.B.C to a model with all main effects and all two-factor interactions improves the fit. (You need to give details of the calculations, the SP for the test, and degrees of freedom of the test statistic used.) [4]

(iii) Suppose you decide to drop the three-factor interaction. Complete the Analysis of Variance table that you would expect to get from GENSTAT, by copying the version below into your answer book and filling in the missing entries. You should explain how you calculated the entries. [5]

```
Source of variation     d.f.       s.s.       m.s.    v.r.
A
B
C
A.B
A.C
B.C
Residual
Total                      15    122.438
```

(c) After some further analysis, further terms were dropped and the following Analysis of Variance table was produced (Model 2).

Model 2

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

Variate: time

Source of variation      d.f.       s.s.       m.s.    v.r.  F pr.
A                           1     14.063     14.063    7.86  0.017
B                           1     18.062     18.062   10.09  0.009
C                           1     18.062     18.062   10.09  0.009
B.C                         1     52.563     52.563   29.37  <.001
Residual                   11     19.687      1.790
Total                      15    122.438
```

(i) Comment on the statistical significance of all terms left in the model. [2]

(ii) Explain how you would use the residual normal probability plot and the plot of the residuals against the fitted values to check the model fit. [2]

(iii) Table 3 gives the fitted values for Model 2.

Table 3 Table of fitted values computed with Model 2

	Factors			Fitted values	
Treatment	A	B	C	$\hat{y}_{i1}$	$\hat{y}_{i2}$
1	1	1	1	10.94	10.94
2	1	1	−1	9.44	9.44
3	1	−1	1	5.19	5.19
4	1	−1	−1	10.94	10.94
5	−1	1	1	9.06	9.06
6	−1	1	−1	7.56	7.56
7	−1	−1	1	3.31	3.31
8	−1	−1	−1	9.06	9.06

On the basis of the tables of means for different factor levels (given in the output from Model 1), explain why the fitted values for treatments 1 and 4 are equal, as well as the fitted values for treatments 5 and 8. [3]

(iv) Use the fitted values to suggest new experimental settings to decrease the reaction time. [2]

(d) Suppose that another factor, D, in addition to A, B and C, had also been thought to affect reaction time, but that the engineers were not allowed (for management reasons) to make more than 16 experimental runs of the reaction. How might they have designed the experiment? What extra assumption would they have had to make in order to analyse the resulting data? [3]

Question 6

In a study investigating the role of age and experience in traffic accidents, the number of accidents in a group of bus drivers over a 5 year period was recorded. The drivers were split into 4 age-groups: 21–30 years, 31–40 years, 41–50 years and 51 years and over, on the basis of their age in the middle of the 5 year study period. Additionally, years of employment as a bus driver, a proxy for how experienced the driver was, was split into 8 groups: 1 year, 2–3 years, 4–6 years, 7–11 years, 12–16 years, 17–21 years, 22–26 years and 27 years or more. As not all bus drivers worked over the entire 5 year period and in order to take more accurate account of the drivers' increasing experience, the investigators were interested in the rate of accidents measured by the number of accidents per person-year (1 person-year is the equivalent of 1 year worked by 1 person).

The data were recorded in a GENSTAT spreadsheet with the following columns:

`accid` : Number of accidents
`peryears` : Number of person-years
`age` : Age (coded as approximate mid-point of group,
i.e. 25.5, 35.5, 45.5 and 58)
`experien` : Years of employment (coded as approximate mid-point of group,
i.e. 1, 2, 5, 9, 14, 19, 24 and 30)

It was decided to model the rate of accidents using Poisson regression with `accid` as the response, the canonical link and including log(`peryears`) as an offset (i.e. including log(`peryears`) in the model with a coefficient of 1).

(a) Describe in detail what assumptions are made with this kind of model. [4]

(b) Explain why this type of model might be better than:

(i) Poisson regression using rate (i.e. `accid/peryears`) as the response; [2]

(ii) Poisson regression using `accid` as the response and simply including log(`peryears`) as one of the explanatory variates. [2]

(c) The GENSTAT system was then used to fit models involving `experien` and `age`. (`lpyears` = log(`peryears`).) The following are extracts of GENSTAT output for two such models, together with some potentially useful χ^2 quantiles.

Model 1

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

  Response variate: accid
     Distribution: Poisson
    Link function: Log
   Offset variate: lpyears
     Fitted terms: Constant + experien + age

*** Summary of analysis ***

                                        mean  deviance
              d.f.     deviance     deviance     ratio
Regression       4        82.18       20.545     20.54
Residual        16        42.75        2.672
Total           20       124.93        6.247

Change          -3       -12.19        4.064      4.06
* MESSAGE: ratios are based on dispersion parameter with value 1
```

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

                                estimate         s.e.     t(*)
Constant                          0.6902       0.0574    12.03
experien                        -0.04005      0.00723    -5.54
age 35.50                        -0.1525       0.0806    -1.89
age 45.50                         -0.165        0.111    -1.49
age 58                             0.235        0.187     1.26
* MESSAGE: s.e.s are based on dispersion parameter with value 1

 CUCHISQU((12.19; 3))
    0.006760
```

Model 2

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

 Response variate: accid
     Distribution: Poisson
    Link function: Log
   Offset variate: lpyears
     Fitted terms: Constant + experien + age + experien.age


*** Summary of analysis ***

                                           mean  deviance
              d.f.     deviance        deviance     ratio
Regression       7        91.34          13.049     13.05
Residual        13        33.59           2.584
Total           20       124.93           6.247

Change          -3        -9.16           3.054      3.05
* MESSAGE: ratios are based on dispersion parameter with value 1


*** Estimates of regression coefficients ***

                                estimate         s.e.     t(*)
Constant                          0.8816       0.0922     9.56
experien                         -0.0965       0.0235    -4.10
age 35.50                         -0.290        0.135    -2.14
age 45.50                         -0.586        0.195    -3.01
age 58                            -0.187        0.416    -0.45
experien.age 35.50                0.0481       0.0267     1.80
experien.age 45.50                0.0750       0.0264     2.84
experien.age 58                   0.0660       0.0285     2.31
* MESSAGE: s.e.s are based on dispersion parameter with value 1

 CUCHISQU((9.16; 3))
     0.02724
```

(i) Was **experien** treated as a variate or a factor in Models 1 and 2? What advantages and disadvantages does this choice have? *[6]*

(ii) From a graph of rate against experience (Figure 2), the investigators stated that for the rate of accidents “age appears to be a modifying factor during the first few years of employment” noting that drivers aged 21–30 had a higher rate of accidents in their first few years of employment compared with drivers aged 31–40 at the same stage in their career. Does the analysis in Models 1 and 2 confirm this view? Briefly explain why or why not. *[5]*

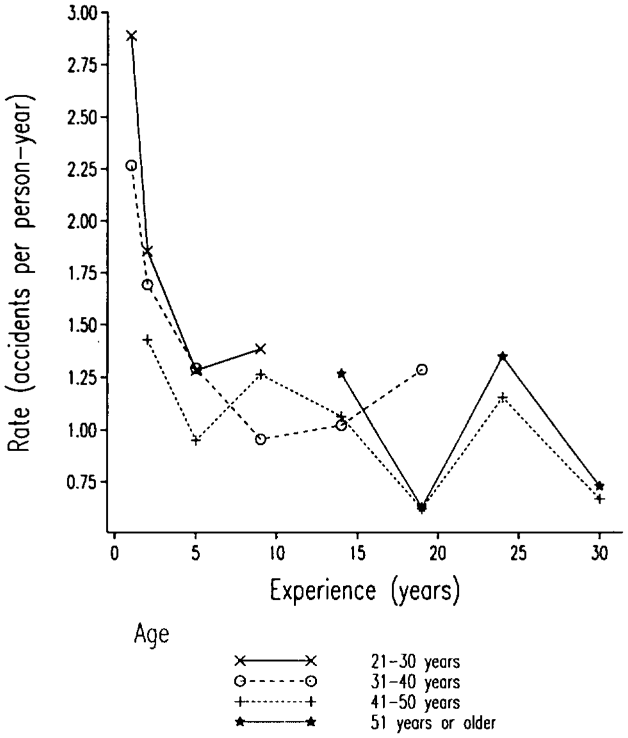


Figure 2

(d) Finally it was decided to fit *lexper* = log(*experien*) rather than experien.

(i) From the output given below (Model 3), would you say that this represents a better model than Model 1 or Model 2? Briefly explain your answer. [2]

(ii) Describe two graphs you would produce to check Model 3 and what you would expect to see in each graph if the model does indeed fit adequately. [4]

Model 3

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

 Response variate: accid
     Distribution: Poisson
    Link function: Log
   Offset variate: lpyears
     Fitted terms: Constant, lexper

*** Summary of analysis ***

                                       mean  deviance
                d.f.    deviance   deviance     ratio
Regression         1       97.90     97.897     97.90
Residual          19       27.04      1.423
Total             20      124.93      6.247

Change            -1      -97.90     97.897     97.90
* MESSAGE: ratios are based on dispersion parameter with value 1

 *** Estimates of regression coefficients ***

                    estimate        s.e.      t(*)
Constant              0.8311      0.0611     13.60
lexper               -0.3199      0.0323     -9.90
* MESSAGE: s.e.s are based on dispersion parameter with value 1
```

Question 7

An educational psychologist was interested in whether there are associations between a high school student's intention to attend a college, parental encouragement, the family's social class and the sex of the student. To investigate the existence of such associations, she interviewed 10,318 Wisconsin high school students and recorded, for each student, the values of the variables sex (male or female); plan (whether the student planned to go to college, recorded as yes or no); support (low support or high support from parents) and class (social class of the family, recorded as low, low/middle, middle/high or high). The data are summarized in the contingency table given below. Each cell of the contingency table contains the frequency of students who gave an answer with the corresponding combination of levels of the 4 factors.

		male		female	
class	plan	low support	high support	low support	high support
low	yes	35	133	31	71
	no	749	233	1078	150
low/middle	yes	38	303	57	210
	no	627	330	798	284
middle/high	yes	37	467	52	344
	no	420	374	611	339
high	yes	26	800	36	736
	no	153	266	217	313

(a) Suppose that you decide to analyse the data with a log-linear model. Indicate the variable that is treated as the response variable and the assumption that you make about the distribution of this variable in fitting a log-linear model. [3]

(b) Describe briefly how you would enter the data in a GENSTAT spreadsheet, to be able to analyse them with a log-linear model in GENSTAT. [3]

(c) Write down the degrees of freedom that are needed to estimate (i) the class main effect, and (ii) the interaction between support and class. [2]

(d) The frequencies in the table above were stored in a variable called count in GENSTAT and they were analysed by using Log-linear modelling in the Analysis field of GENSTAT's Generalized Linear Models dialogue box. Appended below are some Summary of Analysis tables, that were generated fitting particular models, and some relevant χ^2 quantiles.

Model 1

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

 Response variate: Count
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant + Plan + Class +  Sex + Support +
                   Plan.Class + Plan.Support + Class.Support +
                   Plan.Sex + Class.Sex + Support.Sex + Plan.Class.Support +
                   Plan.Class.Sex + Plan.Support.Sex + Class.Support.Sex

 *** Summary of analysis ***
                                         mean     deviance
                d.f.      deviance     deviance     ratio
 Regression      28       7422.950      265.105    265.11
 Residual         3          3.041        1.014
 Total           31       7425.991      239.548

Change          -28      -7422.950      265.105    265.11

 CUCHISQU((3.041; 3))
      0.3853
```

Model 2

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

 Response variate: Count
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant + Plan + Class + Sex + Support + Class.Plan +
                   Class.Sex + Plan.Sex + Class.Support + Plan.Support +
                   Sex.Support + Class.Sex.Support


*** Summary of analysis ***

                                          mean  deviance
               d.f.     deviance      deviance     ratio
 Regression     21      7418.715      353.2721    353.27
 Residual       10         7.276        0.7276
 Total          31      7425.991 239.5481

 Change        -21     -7418.715      353.2721    353.27

 CUCHISQU((7.276; 10))
      0.6992
```

Model 3

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

 Response variate: Count
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant + Plan + Class + Sex + Support + Class.Plan +
                   Class.Sex + Plan.Sex + Class.Support + Plan.Support +
                   Sex.Support


*** Summary of analysis ***

                                           mean  deviance
               d.f.     deviance       deviance     ratio
 Regression      18      7403.43        411.302    411.30
 Residual        13        22.56          1.735
 Total           31      7425.99        239.548

 Change         -18     -7403.43        411.302    411.30

 CUCHISQU((22.56; 13))
     0.04727

 CUCHISQU((15.284; 3))
    0.001589
```

(i) Consider Model 1. Describe the terms that are included in the fitted model. Conduct a formal statistical test to evaluate the goodness of fit of the model. [3]

(ii) Consider Model 2. Conduct a formal statistical test to evaluate the goodness of fit of this model. Compare Model 2 and Model 1 and decide which one is preferable. [5]

(iii) Consider testing whether the class.sex.support interaction should be in the model. What is the value of the test statistic for this test, and which distribution should it be compared with? Use the output for Models 2 and 3 to conduct a formal statistical test of whether the class.sex.support interaction should be in the model. [4]

(iv) Describe in words Model 2. Use the fitted values displayed in Tables 4 and 5 below to describe, in particular, how the plan changes according to the support and the sex of the student, and how parental support differs according to the family's social class and the student's sex. [5]

Table 4

		plan	
support	sex	Yes	No
low	Male	144	1941
	Female	168	2712
high	Male	1695	1211
	Female	1369	1078

Table 5

	class							
	low		low/middle		middle/high		high	
support	male	female	male	female	male	female	male	female
low	784	1109	665	855	457	663	179	253
high	366	221	633	494	841	683	1066	1049

[END OF QUESTION PAPER]