M346 1998 Exam Paper

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

The normal, Bernoulli, binomial, Poisson, t, χ^2 and F distributions all feature strongly in *M346*. Write an essay outlining the main roles of each of these distributions in the general framework of generalized linear modelling.

Your answer should include:

- brief descriptions of the distributions; [6]
- the identification of those distributions that most often act as response distributions, including the breadth of situations in which such response distributions are applicable; [8]
- the roles of the other distributions in the list. [7]

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

Question 2

Describe how the analysis of a multiple linear regression model with errors distributed according to the normal distribution with constant variance fits into the generalized linear model (GLM) framework.

Your answer should include:

- descriptions of the normal multiple linear regression model and the GLM, and the matching up of the two; [6]
- brief descriptions of GENSTAT's 'Summary of Analysis' tables in the two cases, and the matching up of the two (including indicating how tests are performed); [12]
- mention of the residuals used in the normal case, and a brief discussion of whether they match up with GENSTAT's default residuals in the GLM case. [3]

Do not concern yourself with the selection of variables or further diagnostics in either case.

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

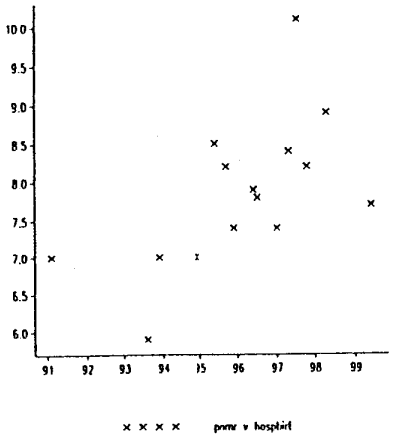
Data for the year 1990 for each of the fourteen health regions into which England was then divided, and for Wales, were obtained from official UK Government publications. They were used as part of an investigation of the causes of infant mortality. The data give, for each of the fifteen regions involved, the percentage of births that took place in National Health Service hospitals where consultant obstetricians were available (variable `hospbirt`), and the perinatal mortality rate, expressed as the number of stillbirths and deaths in the first week of life, per thousand total births (variable `pnmr`). The data are given in Table 1 (in the order they appeared when input into GENSTAT).

Table 1

Region	hospbirt	pnmr
Wales	97.0	7.4
Northern	96.5	7.8
Yorkshire	96.4	7.9
Trent	97.3	8.4
East Anglian	93.6	5.9
North West Thames	95.9	7.4
North East Thames	95.4	8.5
South East Thames	97.8	8.2
South West Thames	93.9	7.0
Wessex	91.1	7.0
Oxford	95.7	8.2
South Western	94.9	7.0
West Midlands	97.5	10.1
Mersey	99.4	7.7
North Western	98.3	8.9

In the following analysis, `pnmr` is treated as the response variable.

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



On the basis of this plot, would you say it is reasonable to fit a simple linear regression model to the data? Briefly explain why or why not. Would you consider transforming the data before fitting the model? Again, explain why or why not. [5]

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

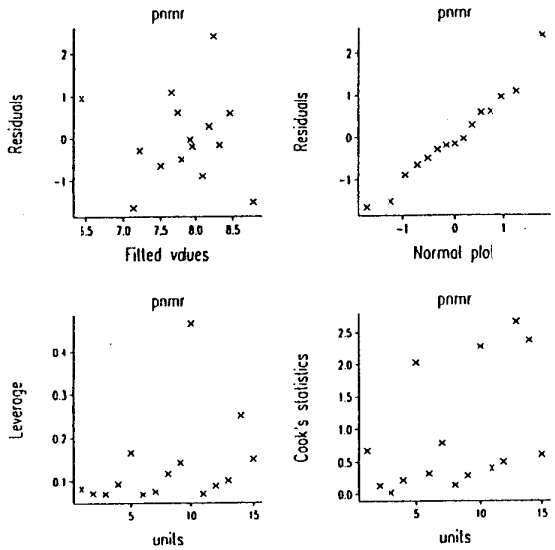
```
***** Regression Analysis *****

 Response variate: pnmr
     Fitted terms: Constant, hospbirt


*** Summary of analysis ***

             d.f.        s.s.        m.s.      v.r.  F pr.
Regression      1       4.887      4.8865      7.35  0.018
Residual       13       8.643      0.6648
Total          14      13.529      0.9664

Change         -1      -4.887      4.8865      7.35  0.018

Percentage variance accounted for 31.2
Standard error of observations is estimated to be 0.815
* MESSAGE: The following units have large standardized residuals:
                13        2.41
* MESSAGE: The following units have high leverage:
                10        0.47


*** Estimates of regression coefficients ***

             estimate        s.e.      t(13)  t pr.
Constant        -19.3        10.0      -1.93  0.075
hospbirt        0.282       0.104       2.71  0.018
```

(i) What is the estimated regression equation resulting from this analysis? [1]

(ii) The output gives the required information to test the hypothesis that the slope of the regression line is zero. Report what this information is together with the results of this test, giving your conclusions clearly. Does this test establish whether changing the percentage of hospital births in a region will cause a change in the perinatal mortality rate? Briefly explain why or why not. [5]

(iii) Without going into details about the formulas involved, explain what the purpose is of GENSTAT producing warning messages about units with large standardized residuals and high leverage in analyses like this. [5]

(c) The collection of figures below comprises a plot of standardized residuals against fitted values, a normal probability plot of the standardized residuals, and index plots of leverages and Cook statistics, for the model fitted in part (b).



(i) Is there any feature of these plots that gives an indication that the assumptions of the simple linear regression model do not hold for these data? [2]

(ii) The leverage plot shows one point with a considerably higher leverage than all the others. However, the plot of Cook statistics shows that more than one point has a relatively high Cook statistic. Briefly explain why there are points that have relatively low leverage but high Cook statistics. [3]

(d) Describe two ways in which you could continue with your analysis of these data. [4]

Question 4

In the manufacture of a dyestuff, one characteristic of interest in the finished product is its *tinctorial strength*. The tinctorial strength is affected by the presence of an impurity that is difficult to remove from the raw materials used to make the dyestuff. There are two different processes that can be used to manufacture the dyestuff, process A and process B. An experiment was carried out to determine whether the tinctorial strength of the dyestuff produced by the two processes was different.

The results of the experiment were stored in a GENSTAT spreadsheet containing three columns:

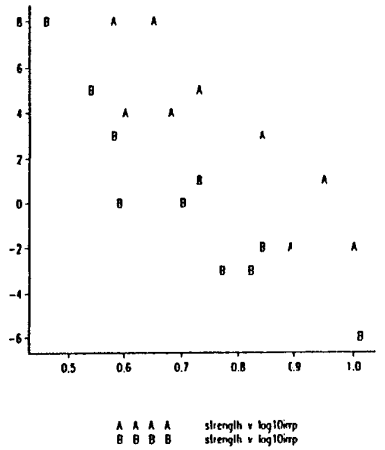
strength the tinctorial strength of the manufactured dyestuff relative to a standard;

log10imp the logarithm (base 10) of the percentage impurity in the raw materials;

process a factor indicating the process (A or B) used.

There were ten production runs for each process, giving a total of twenty rows in the spreadsheet. Each production run used the whole of a single batch of raw materials, so a total of twenty batches was used; these batches were randomly allocated to the production runs.

(a) Below is a scatterplot of strength against log10imp, with plotting symbols corresponding to the process used.



(i) What does the scatterplot indicate about the relationship between strength and the other two variables, log10imp and process? [3]

(ii) Based on your inspection of the scatterplot and the information given about the experiment, would it be reasonable and helpful to fit a single line to all the data, ignoring which production process was used? [3]

(iii) For what reason did the experimenters transform the impurity values, by taking logarithms? [1]

(b) The GENSTAT menu system was used to carry out a *simple linear regression with groups*, which produced the follcwing output.

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

 Response variate: strength
     Fitted terms: Constant + log10imp


*** Summary of analysis ***

             d.f.        s.s.        m.s.       v.r.  F pr.
Regression      1        171.        171.      22.80  <.001
Residual       18        135.          8.
Total          19        307.         15.

Change         -1        -171.       171.      22.80  <.001

Percentage variance accounted for 53.4
Standard error of observations is estimated to be 2.74
* MESSAGE: The following units have high leverage:
              8        0.20
             12        0.21
             18        0.22


*** Estimates of regression coefficients ***

                         estimate        s.e.     t(18)  t pr.
Constant                    15.84        3.03      5.22  <.001
log10imp                   -19.32        4.05     -4.78  <.001


19.....


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

 Response variate: strength
     Fitted terms: Constant + log10imp + process


*** Summary of analysis ***

             d.f.        s.s.        m.s.       v.r.  F pr.
Regression      2        250.        125.      37.33  <.001
Residual       17         57.          3.
Total          19        307.         16.

Change         -1        -78.         78.      23.44  <.001

Percentage variance accounted for 79.3
Standard error of observations is estimated to be 1.83
* MESSAGE: The following units have high leverage:
             13        0.11
```

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

                    estimate        s.e.     t(17)  t pr.
Constant               19.83        2.19      9.07  <.001
log10imp              -22.00        2.76     -7.99  <.001
process B              -4.042       0.835    -4.84  <.001
```

21.....

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

 Response variate: strength
     Fitted terms: Constant + log10imp + process + log10imp.process

*** Summary of analysis ***

             d.f.         s.s.         m.s.      v.r.  F pr.
Regression      3         251.           84.     24.08  <.001
Residual       16          56.            3.
Total          19         307.           16.

Change         -1          -1.            1.      0.37  0.554

Percentage variance accounted for 78.5
Standard error of observations is estimated to be 1.86
* MESSAGE: The following units have high leverage:
                  18        0.48

*** Estimates of regression coefficients ***

                    estimate        s.e.     t(16)  t pr.
Constant               18.37        3.28      5.60  <.001
log10imp              -20.10        4.22     -4.76  <.001
process B              -1.52        4.26     -0.36  0.726
log10imp.process B     -3.42        5.65     -0.60  0.554
```

(i) Three models have been fitted. Explain what each of these models implies about the relationship between strength, log10imp and process. Why are these three models the natural ones to consider here? [5]

(ii) Use the GENSTAT output to decide which of the three models is the most appropriate for these data. Explain how you have reached your decision. [3]

(iii) Using the model that you chose as most appropriate in part (b)(ii), calculate an estimate of the expected value of strength for dyestuff made using process *B* and raw materials with a log10imp value of 0.7. [2]

(c) In each of the tables entitled `*** Estimates of regression coefficients ***` for the second and third models fitted, you will find a row labelled process B. The significance probability in the second model's table is given as `<.001` and that in the third model's is `0.726`. Explain the difference between the tests corresponding to these two significance probabilities and, hence, why it is reasonable for these significance probabilities to be so different. [3]

(d) If you were to split the dataset into two and perform simple linear regressions on each part of the dataset separately, would you be fitting the same model as GENSTAT did in its third run above? Give a reason for your answer. [2]

(e) Suppose that the batches of raw material were large enough for two production runs to be carried out using a single batch. In this case, how could the experiment have been done differently in a way that still allowed the processes to be compared, but without the need to measure the impurity of each batch? [3]

Question 5

From 1977–1985, Dr. Brian Kerry and Dr. David Crump of the Department of Entomology and Nematology at IACR–Rothamsted carried out an experiment to study the effect different agricultural practices had upon the yield of cereals infected with cyst nematodes. Cyst nematodes attack the crops and reduce the crop yield.

In all, 64 plastic bins were used in this experiment. Each bin was filled with one of four types of soil, *A*, *B*, *C* and *D*; there were sixteen bins for each type of soil. All the bins were (deliberately) infected with cyst nematodes at the start of the experiment.

The bins were formed into sixteen groups each of four bins. Each group contained one bin for each of the four types of soil. The groups of bins were placed in a raised gravel bed and arranged in four blocks each containing four groups of bins. Within each group there were four places, one for each bin, and the four bins were randomly assigned to the four places. Similarly, the positions of the groups within each block were chosen randomly. [Hint: you may find it helpful to sketch a diagram for yourself.]

Within each block, each group received a different one of four treatment combinations. The four treatment combinations had a 2×2 factorial structure. The two factors were: *sterilising agent*, which was either *formalin* or *control*; *cultivar*, which was either *Tyra* or *Triumph*.

If the sterilising agent was formalin, then the group of bins was sprayed with formalin during the period 1977–1980; if it was control then no formalin was applied. Formalin is a fungicide, which is volatile and difficult to apply to small areas. Applying formalin should have increased the cyst nematode population, because the fungus that is killed by the formalin would normally inhibit the growth of the nematode population.

If the cultivar was Tyra, then the soil in each group was sown with a type of spring barley called 'Tyra' for the years 1981–1985; similarly, for Triumph, a different spring barley called 'Triumph' was sown. The Tyra variety is resistant to attack by cyst nematodes, whereas Triumph is susceptible to cyst nematode attack.

The data collected from this experiment in 1982 have been stored in a GENSTAT spreadsheet with 64 rows and 6 columns. The columns are:

bloct which of the four blocks the bin is in;

plot which group the bin is in, within its block;

sterlnt the sterilising agent applied to the group the bin is in;

cultivar the cultivar applied to the group the bin is in;

soil the type of soil in the bin;

yield the yield in grams of barley from the bin.

(a) (i) What is the general name for this type of experimental design; that is a design where the experimental units are placed in groups and one or more treatments are applied to the groups? What is the name of the design that has been used for the groups? [2]

(ii) What is the usual reason for applying a treatment to groups of experimental units, as cultivar and sterilising agent were in this experiment? What is the specific reason, given in the description of the experiment, for applying sterilising agent to groups in this experiment? [2]

(iii) Having selected the appropriate type of design in GENSTAT's Analysis of Variance dialogue box, you would have to fill in four fields to specify the model to fit. The four fields are: Y-Variate; Treatment Structure; Blocks; Whole Plots.

EITHER write down what you would have to enter in these four fields,

OR write down what you would have to enter in the Y-Variate, Treatment Structure and Block Structure fields, if you had selected General Analysis of Variance for the Design field of the dialogue box. [4]

(b) The output from fitting an appropriate model to these data is given below.

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

Variate: yield

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

block stratum               3    1057.4     352.5      2.49

block.plot stratum
cultivar                    1    1279.9    1279.9      9.04  0.015
sterlant                    1    1328.6    1328.6      9.38  0.014
cultivar.sterlant           1     201.6     201.6      1.42  0.263
Residual                    9    1274.7     141.6      0.41

block.plot.*Units* stratum
soil                        3    1770.6     590.2      1.72  0.180
cultivar.soil               3     996.8     332.3      0.97  0.417
sterlant.soil               3    1939.8     646.6      1.89  0.149
cultivar.sterlant.soil      3     280.6      93.5      0.27  0.844
Residual                   36   12328.1     342.4

Total                      63   22458.3


* MESSAGE: the following units have large residuals.

block IV   plot iv          -9.8   s.e. 4.5

block I    plot i    *units* 1          -44.7   s.e. 13.9
block IV   plot iv   *units* 2           47.8   s.e. 13.9

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

Variate: yield

Grand mean  63.6

 cultivar      Tyra  Triumph
               59.1     68.0

 sterlant   control formalin
               68.1     59.0

     soil         A        B        C        D
               70.2     62.7     65.6     55.7
```

cultivar	sterlant	control	formalin
Tyra		65.4	52.8
Triumph		70.8	65.3

cultivar	soil	A	B	C	D
Tyra		70.2	53.8	57.8	54.6
Triumph		70.2	71.6	73.5	56.8

sterlant	soil	A	B	C	D
control		68.1	70.0	66.6	67.8
formalin		72.3	55.5	64.6	43.6

cultivar	sterlant	soil	A	B	C	D
Tyra	control		69.0	65.3	61.9	65.5
	formalin		71.4	42.4	53.6	43.7
Triumph	control		67.2	74.7	71.3	70.1
	formalin		73.2	68.6	75.7	43.5

*** Standard errors of differences of means ***

Table	cultivar	sterlant	soil	cultivar sterlant
rep.	32	32	16	16
d.f.	9	9	36	9
s.e.d.	2.98	2.98	6.54	4.21

Table	cultivar soil	sterlant soil	cultivar sterlant soil
rep.	8	8	4
s.e.d.	8.55	8.55	12.09
d.f.	43.32	43.32	43.32

Except when comparing means with the same level(s) of

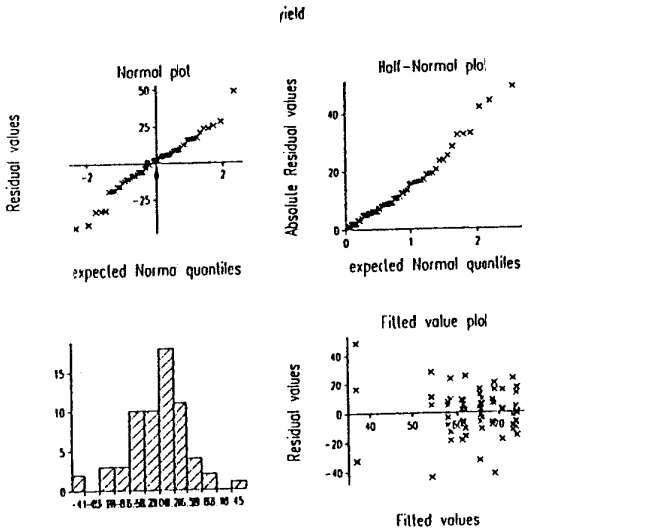
cultivar	9.25		
d.f.	36		
sterlant		9.25	
d.f.		36	
cultivar.sterlant			13.09
d.f.			36

(i) Use this output to select the treatment terms that should be retained in a simplified, but adequate, model for these data. Identify the parts of the output that you have used in making your model selection. [3]

(ii) Write a brief summary of your conclusions for the experimenters. You should identify which parts, if any, of the output you would present to the experimenters; if there is other information, not included in the output, that you would have liked to include with your summary mention that too. Is there anything surprising about the results of the analysis? [6]

(iii) In the ANOVA table in the output, why is there no entry in the block stratum row for F pr.? [2]

(c) Below is a set of diagrams that was produced after fitting a simplified model to these data.



(i) What, in general, are the sort of diagrams above used for? [1]

(ii) What do the particular diagrams above tell you about the model that has been fitted? [5]

Question 6

In an experiment reported by A.L. Strand in 1930, the aim was to determine the relationship between concentration of an insecticide, carbon disulphide, and the proportion of flour beetles, *Tribolium confusum*, killed by five hours of exposure to the insecticide. Sixteen batches, each of *about* thirty beetles, were exposed to the carbon disulphide. The number of beetles killed in each batch was recorded. Eight different concentrations of carbon disulphide were used in the experiment and two batches of beetles were used for each different concentration.

The data from this experiment have been stored in a GENSTAT spreadsheet with sixteen rows, one row per batch, and three columns. The columns were:

cs2 the concentration of carbon disulphide gas, measured in mg/l;

y the number of beetles that had been killed after five hours of exposure;

n the total *number of* beetles in the batch.

(a) A model was fitted to these data using GENSTAT; the output produced is given below.

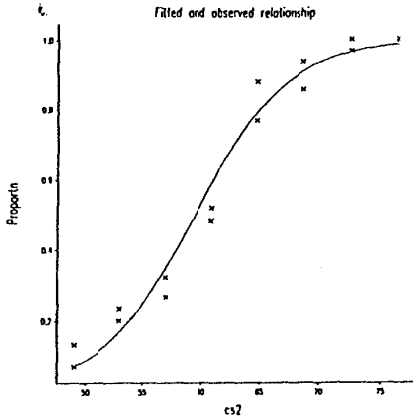
```
***** Regression Analysis *****

 Response variate: y
  Binomial totals: n
     Distribution: Binomial
    Link function: Logit
     Fitted terms: Constant, cs2


*** Summary of analysis ***

                                      mean  deviance
             d.f.     deviance     deviance     ratio
Regression      1          277.         277.    276.64
Residual       14           13.          1.
Total          15          289.         19.

Change         -1         -277.         277.    276.64
* MESSAGE: ratios are based on dispersion parameter with value 1


* MESSAGE: The residuals do not appear to be random;
           for example, fitted values in the range 1.E+01 to 3.E+01
           are consistently larger  than observed values
           and fitted values in the range 3.E+01 to 3.E+01
           are consistently smaller than observed values


*** Estimates of regression coefficients ***

                          estimate         s.e.      t(*)
Constant                   -14.81         1.29    -11.49
cs2                         0.2492       0.0214     11.66
* MESSAGE: s.e.s are based on dispersion parameter with value 1
```

(i) Write down the model that has been fitted, including the distributional assumptions made. [3]

(ii) What does the output tell you about the relationship between the concentration of carbon disulphide and the proportion of beetles killed? [3]

(iii) Use the fitted model to estimate the proportion of beetles that would be killed if they were exposed to a concentration of 60 mg/l. [2]

(iv) Below is the fitted model plot for this model.



What characteristics of this plot indicate that the fitted model is inadequate? [1]

(b) Having decided that the model fitted in part (a) was inadequate, a statistician decided to fit a more complicated model. The output produced by fitting this model is given below.

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

 Response variate: y
  Binomial totals: n
     Distribution: Binomial
    Link function: Logit
     Fitted terms: Constant + cs2
        Submodels: PCL(cs2; 2)


*** Summary of analysis ***

                                      mean  deviance
             d.f.     deviance     deviance     ratio
Regression      2          281.         141.    140.61
Residual       13            8.          1.
Total          15          289.         19.

Change         -2         -281.         141.    140.61
* MESSAGE: ratios are based on dispersion parameter with value 1


*** Estimates of regression coefficients ***

                          estimate         s.e.      t(*)
Constant                       8.0         11.0      0.72
cs2 Lin                     -0.517        0.373     -1.38
cs2 Quad                   0.00637      0.00314      2.03
* MESSAGE: s.e.s are based on dispersion parameter with value 1
```

(i) How does this new model differ from the model fitted in part (a)? Does the estimated coefficient for cs2 lin mean that the proportion decreases with increased level of insecticide? [2]

(ii) The statistician decided that this more complicated model was better than that fitted in part (a). Explain how the statistician came to this conclusion. [5]

(c) To conclude the model selection part of the analysis, our statistician went on to create an eight-level factor representing the eight distinct concentrations of carbon disulphide and added this factor to the model of part (b). The factor was named fcs2. The output from adding this factor to the model is given below.

```
* MESSAGE: Term fcs2 cannot be fully included in the model
  because 2 parameters are aliased with terms already in the model

  (fcs2 72.61) = - 40.57 + (cs2 Lin)*1.357 - (cs2 Quad)*0.01080 -
    (fcs2 52.99) - (fcs2 56.91)*1.665 - (fcs2 60.84)*1.998 -
    (fcs2 64.76)*1.998 - (fcs2 68.69)*1.665

  (fcs2 76.54) = 32.98 - (cs2 Lin)*1.127 + (cs2 Quad)*0.009260 +
    (fcs2 52.99)*0.7140 + (fcs2 56.91)*1.141 + (fcs2 60.84)*1.284 +
    (fcs2 64.76)*1.141 + (fcs2 68.69)*0.7125

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

 Response variate: y
  Binomial totals: n
     Distribution: Binomial
    Link function: Logit
     Fitted terms: Constant + cs2 + fcs2
        Submodels: POL(cs2; 2)


*** Summary of analysis ***

                                        mean  deviance
              d.f.     deviance     deviance     ratio
Regression       7         284.          41.     40.60
Residual         8           5.           1.
Total           15         289.          19.

Change          -7        -284.          41.     40.60
* MESSAGE: ratios are based on dispersion parameter with value 1


*** Estimates of regression coefficients ***

                          estimate         s.e.      t(*)
Constant                     249.         1865.      0.13
cs2 Lin                     -8.8           63.7     -0.14
cs2 Quad                   0.074          0.524      0.14
fcs2 52.99                   5.6           40.4      0.14
fcs2 56.91                   8.3           64.5      0.13
fcs2 60.84                   9.3           72.6      0.13
fcs2 64.76                   8.7           64.6      0.13
fcs2 68.59                   4.8           40.3      0.12
fcs2 72.61                     0              *         *
fcs2 76.54                     0              *         *
* MESSAGE: s.e.s are based on dispersion parameter with value 1
```


(i) What part of this output indicates that there is nothing to be gained by adding more terms in cs2 (a cubic term, for instance) to the model fitted in part (b)? [2]

(ii) Why are only 5 degrees of freecom used up by adding fcs2 to the model, rather than the 7 degrees of freedom that would normally be needed for an eight-level factor? [3]

(d) Having settled on the model in part (b), the statistician went on to consider whether the assumptions of the model were satisfied. The four residual plots offered by the Model Checking dialogue box of GENSTAT are: Fitted Values; Index; Half-Normal; Normal.

For each of these four types of diagram, say whether you would plot them. For the diagrams that you would plot, indicate what characteristics you would expect to see if the model assumptions held. For the diagrams that you would NOT plot, say why these diagrams are not useful. [4]

Question 7

Table 2 shows data on consumer preferences for competing detergent brands X and M in a trial in which each consumer compared the two detergents, but was not told which detergent was which. Altogether, some 1008 consumers were involved in the trial, and as well as recording which of brands X and M they preferred (`pref`, the response variable), each recorded the softness of their water (`water`, at three levels, `soft`, `medium` or `hard`), whether or not they were a previous user of brand M (M was the existing brand, X being a new product), and whether they washed at high or low temperature (`temp`). Each consumer, therefore, contributes one to the count in one of the cells of Table 2, the total number in all the cells adding to 1008.

Table 2

		Previous user of M		Previous nonuser of M	
Water softness	Brand preference	High temperature	Low temperature	High temperature	Low temperature
soft	X	19	57	29	63
	M	29	49	27	53
medium	X	23	47	33	65
	M	47	55	23	50
hard	X	24	37	42	68
	M	43	52	30	42

[Source: *P.N. Ries & H. Smith (1963) The use of chi-square for preference testing in multidimensional problems. Chemical Engineering Progress*, 59, 39–43.]

The data were then analysed using Log-linear modelling in the `Analysis` field of GENSTAT's `Generalized Linear Models` dialogue box. Appended below are the `Summary of analysis` tables arising from the fit of three particular models, along with two useful χ^2 quantiles.

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

 Response variate: count
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant + pref + previous + temp + water +
                   pref.previous + pref.temp + previous.temp +
                   pref.water + previous.water + temp.water +
                   pref.previous.temp + pref.previous.water +
                   pref.temp.water + previous.temp.water


*** Summary of analysis ***

                                    mean  deviance
             d.f.     deviance     deviance     ratio
Regression     21     117.8896       5.6138      5.61
Residual        2       0.7373       0.3687
Total          23     118.6269       5.1577

Change        -21    -117.8896       5.6138      5.61
* MESSAGE: ratios are based on dispersion parameter with value 1
```

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

 Response variate: count
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant + pref + previous + temp + water +
                   pref.previous + pref.temp + previous.temp +
                   pref.water + previous.water + temp.water


*** Summary of analysis ***

                                    mean  deviance
             d.f.     deviance     deviance     ratio
Regression     14     108.781       7.770       7.77
Residual        9       9.846       1.094
Total          23     118.627       5.158

Change        -14    -108.781       7.770       7.77
* MESSAGE: ratios are based on dispersion parameter with value 1


* MESSAGE: The following units have large standardized residuals:
                2        2.40
                4       -2.16


 CUCHISQU((9.846; 9))
      0.3631
```

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

 Response variate: count
     Distribution: Poisson
    Link function: Log
     Fitted terms: Constant + pref + previous + temp + water


*** Summary of analysis ***

                                    mean  deviance
             d.f.     deviance     deviance     ratio
Regression      5      75.70       15.140      15.14
Residual       18      42.93        2.385
Total          23     118.63        5.158

Change         -5     -75.70       15.140      15.14
* MESSAGE: ratios are based on dispersion parameter with value 1


* MESSAGE: The following units have large standardized residuals:
                1       -2.17
               13        3.23
               15       -2.00
               18       -2.54
               21        2.63
               24       -2.26


 CUCHISQU((42.93; 18))
   0.0008187
```

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

(b) What would be the residual deviance associated with using `water*temp*previous*pref` in the `Model to be Fitted` field, and why? [2]

(c) Consider testing whether the `water.temp.previous.pref` 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? Why does the degrees of freedom take the value it does? The *SP* turns out to be 0.6917. What do you conclude? [4]

(d) Which is the highest level of interaction that you need to retain in the model for these data? [5]

(e) Very briefly, what strategy *might* you employ to refine the model further from that decided on *in* part (d)? (You need not give any GENSTAT commands to *perform* this.) [2]

(f) The `Summary of analysis` table corresponding to a refined model GENSTAT found after further analysis of the sort that you should have alluded to in part (e) is given below. Also given are certain tables of totals.

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

 Response variate: count
     Distribution: Poisson
    Link function: log
     Fitted terms: Constant + pref + previous + temp + water +
                   pref.previous + pref.temp + temp.water


*** Summary of analysis ***

                                      mean  deviance
             d.f.     deviance    deviance     ratio
Regression      9       106.74    11.8600     11.86
Residual       14        11.89     0.8490
Total          23       118.63     5.1577

Change         -9      -106.74    11.8600     11.86
* MESSAGE: ratios are based on dispersion parameter with value 1


* MESSAGE: The following units have large standardized residuals:
                13        2.03
```



```
                Total
temp             low        high
pref
   X           338.0       170.0
   M           301.0       199.0


                Total
previous        user     nonuser
pref
   X           207.0       301.0
   M           275.0       225.0


                Total
water           soft      medium        hard
pref
   X           168.0       169.0       171.0
   M           158.0       175.0       167.0
```

Describe in words the selected model. What does the response variable `pref` depend on according to the fitted model? Using the tables given, describe the main features of this dependence. [7]

(g) Had logistic regression been used to regress `pref` on the other explanatory variables, would you have expected the logistic regression and loglinear models to be in exact correspondence? Give a reason for your answer. [2]

[END OF QUESTION PAPER]