UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B. Sc. Examination 2002

STATISTICS

ST52011A (ST230) Statistical Modelling

Duration: 3 hours

Date and time:

Answer four questions. You must answer at least ONE question in each section.

There are 100 marks available on this paper. Each question is worth 25 marks.

Electronic calculators may be used. The make and model should be specified on the script. The calculator must not be programmed prior to the examination. Calculators which display graphics, text or algebraic equations are not allowed.

NOTE: Full details of all calculations are to be shown; pre-programmed statistical tests and procedures on a calculator, apart from mean and standard deviation, must not be used.

WHITE, YEATS & SKIPWORTH: Tables for Statisticians to be provided.

A formula sheet is attached at the end of this paper.

THIS EXAMINATION PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Section A

Question 1 You must answer this question. Three different machines produce a monofilament fibre for a textile company. The process engineer is interested in determining if there is a difference in the breaking strength of the fibre produced by the three machines. However, the strength of fibre is related to its diameter, with thicker fibres being generally stronger than thinner ones. A random sample of five fibre specimens is selected from each machine. The fibre strength (y) in pounds and corresponding diameter (x) in 10^{-3} inches for each specimen is shown in the table below.

Machine 1		Machine 2		Machine 3	
y	x	y	x	y	x
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15

(a) Look at the MINITAB output and graphs entitled Analysis 1.

(i) The data have been analysed as a one way model, ignoring the x variable. Explain what is meant by a one way model and the necessary assumptions. [2]

(ii) What conclusion do you draw from the ANOVA table? [2]

(iii) Comment on the plot of residuals versus fitted values. [2]

(iv) Explain why the Tukey procedure is used when there are no pre-planned comparisons. What are the conclusions from this procedure in this case? [3]

(v) Comment on the plot of residuals versus x. [3]

```
Analysis 1

One-way ANOVA: y versus machine

Analysis of Variance for y
Source     DF        SS        MS        F        P
machine     2     140.4      70.2     4.09    0.044
Error      12     206.0      17.2
Total      14     346.4
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  -------+---------+---------+---------
1           5    41.400     4.827              (-------*-------)
2           5    43.200     3.701                 (-------*-------)
3           5    36.000     3.808   (-------*-------)
                                   -------+---------+---------+---------
Pooled StDev =    4.143                35.0      40.0      45.0

Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0206

Critical value = 3.77

Intervals for (column level mean) - (row level mean)

                 1           2

       2      -8.786
               5.186

       3      -1.586       0.214
              12.386      14.186
```

(b) Look at the MINITAB output and graph entitled Analysis 2.

(i) Explain which two models are being fitted to the data now. [4]

(ii) What conclusions do you draw from the ANOVA tables regarding the possible parallelism and coincidence of the three regression lines? [4]

(iii) Use the second ANOVA to give a 90% confidence interval for the difference between the intercepts for machines 2 and 3. [3]

(iv) Predict the breaking strength Y of a new fibre specimen of diameter $x = 30 \times 10^{-3}$ inches from machine 1. [2]

```
Analysis 2

Plot y * x

General Linear Model: y versus machine

Factor     Type Levels Values
machine   fixed      3 1 2 3

Analysis of Variance for y, using Adjusted SS for Tests

Source      DF     Seq SS     Adj SS     Adj MS       F      P
machine      2    140.400      2.664      1.332    0.47  0.637
x            1    178.014    171.119    171.119   61.00  0.000
machine*x    2      2.737      2.737      1.369    0.49  0.629
Error        9     25.249     25.249      2.805
Total       14    346.400

Term           Coef   SE Coef        T      P
Constant     17.388     2.959     5.88  0.000
machine
1            -3.816     4.109    -0.93  0.377
2             3.526     4.498     0.78  0.453
x            0.9419    0.1206     7.81  0.000
x*machine
  1          0.1624    0.1645     0.99  0.349
  2         -0.0847    0.1768    -0.48  0.643

Unusual Observations for y

Obs         y       Fit      SE Fit  Residual   St Resid
  7   48.0000   44.9143      0.8726    3.0857      2.16R

R denotes an observation with a large standardized residual.



General Linear Model: y versus machine


Factor     Type Levels Values
machine   fixed      3 1 2 3
```

```
Analysis of Variance for y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
machine     2     140.40      13.28       6.64    2.61  0.118
x           1     178.01     178.01     178.01   69.97  0.000
Error      11      27.99      27.99       2.54
Total      14     346.40

Term          Coef   SE Coef        T      P
Constant    17.177     2.783     6.17  0.000
machine
1           0.1824    0.5950     0.31  0.765
2           1.2192    0.6201     1.97  0.075
x           0.9540    0.1140     8.36  0.000

Unusual Observations for y

Obs         y       Fit      SE Fit  Residual   St Resid
  7   48.0000   45.1080      0.7489    2.8920      2.05R

R denotes an observation with a large standardized residual.
```

Section B

You must answer at least one question from this section.

Question 2 A plant distills liquid air to produce oxygen, nitrogen and argon. The percentage of impurity in the oxygen is thought to be linearly related to the amount of impurities in the air, as measured by the "pollution count" in parts per million (ppm). The following data were collected on 15 successive days.

Purity (%) y	93.3	92.0	92.4	91.7	94.0	94.6	93.6	93.1
Pollution count (ppm) x	1.10	1.45	1.36	1.59	1.08	0.75	1.20	0.99

Purity (%) y	93.2	92.9	92.2	91.3	90.1	91.6	91.9
Pollution count (ppm) x	0.83	1.22	1.47	1.81	2.03	1.75	1.68

(a) Fit a linear regression to the data using $\sum x_i = 20.31$, $\sum y_i = 1387.1$, $\sum x_i^2 = 29.4593$, $\sum y_i^2 = 128436.63$, $\sum x_i y_i = 1873.532$. [6]

(b) What residuals plots would you make to check your model and why? [6]

(c) Calculate s^2, the estimate of the error variance. [4]

(d) Find a 95% confidence interval for the slope of the regression equation. [4]

(e) Find a 90% confidence interval for the mean purity on a day when the pollution count is 1.00. [5]

Question 3 Data were collected in a study of how three distillation properties (X1, X2 and X3) of crude oils, together with the volatility (X4) of the petrol produced (measured by the ASTM end point in degrees F), affect the percentage yield of petrol (Y). Thirty one sets of measurements were made. The following table gives the residual sums of squares (RSS) for all possible multiple regression models using the four explanatory variables.

Model	RSS	Model	RSS
Null	3564.1	X2+X3	3016.6
X1	3347.8	X2+X4	369.9
X2	3038.3	X3+X4	170.6
X3	3210.4	X1+X2+X3	3008.8
X4	1759.7	X1+X2+X4	265.5
X1+X2	3038.0	X1+X3+X4	146.0
X1+X3	3205.7	X2+X3+X4	160.6
X1+X4	861.9	X1+X2+X3+X4	134.8

(a) Using a 5% significance level show that backwards elimination and forwards fitting results in the same model being chosen. [12]

(b) List two advantages and two disadvantages of either of these methods of choosing a regression equation. [4]

(c) The following table shows some statistics which are commonly used to help in the choice of a multiple regression model.

```
RSq   RSq(adj)   Cp      S         Model
50.6    49.0   324.5   7.6588       X4
95.2    94.9     8.2   2.4255     X3+X4
95.9    95.5     5.2   2.2835    X1+X3+X4
```

(i) Explain how to calculate RSq (R^2) using the table of Residual Sums of Squares above. What is its disadvantage in this context? [3]

(ii) In what way is RSq(adj) (adjusted R^2) an improvement over R^2? [2]

(iii) How is Mallows' C_p calculated? How is it used as a statistic to aid model choice? [4]

Question 4 The maximum output voltage of a particular type of battery is thought to be influenced by the material used in the plates and the temperature (measured in degrees F) in the location at which the battery is installed. Four replicates of a factorial experiment are run in the laboratory for three materials and three temperatures. The results are shown in the table below.

	Temperature						
Material	50		65		80		Total
1	130	155	34	40	20	70	998
	74	180	80	75	82	58	
2	150	188	136	122	25	70	1300
	159	126	106	115	58	45	
3	138	110	174	120	96	104	1501
	168	160	150	139	82	60	
Total	1738		1291		770		3799

The sum of squares of all the observations $\sum_i \sum_j \sum_k y_{ijk}^2 = 478547$.

(a) Write down the full linear model for such a study, stating clearly the assumptions underlying the model. [5]

(b) Give the full analysis of variance table and perform any necessary tests. Interpret your results. Which model is most suitable for these data? [12]

(c) What is the fitted value for material type 2 at 65 degrees? [2]

(d) Sketch the interaction plot and comment on whether this confirms the results you have found. [6]

Section C

You must answer at least one question from this section.

Question 5 A random sample $y_1, \ldots, y_n$ is drawn from a Negative Binomial distribution $NB(k, p)$ where $k > 0$ is known.

(a) Find and sketch the likelihood function $L(p)$. [4]

(b) Hence show that the loglikelihood function is given by

$$l(p) = nk \log(1 - p) + n\bar{y} \log p,$$

except for an additive constant. [1]

(c) Show that the maximum likelihood estimate (MLE) of p is given by

$$\hat{p} = \frac{\bar{y}}{k + \bar{y}}.$$

[2]

(d) Find the information function $I(p)$ and hence a large sample approximate 95% confidence interval for p. [7]

(e) Given that the loglikelihood of the saturated model is

$$l(p_1, \ldots, p_n) = \sum_{i=1}^{n} k \log(1 - p_i) + \sum_{i=1}^{n} y_i \log p_i,$$

except for the same additive constant, find its maximised value l_{sat}. [6]

(f) Hence find the deviance of the null model with Negative Binomial errors and state its large sample distribution (when both n and k are large). [5]

Question 6 120 patients attending a general practitioner's surgery were given a general health questionnaire on which they scored between 0 and 12 (GHQ). Each patient was subsequently given a full psychiatric examination by a psychiatrist who did not know the GHQ score, and patients were classified either as cases (requiring psychiatric treatment) or as non-cases. The number of cases Y and number of patients K was recorded for each of 17 attained GHQ scores, and for males (SEX=1) and females (SEX=2) separately.

A plot of the proportion of cases and their empirical logits showed a rapid change from non-cases to cases between GHQ 1 and 4 for males, and a similarly rapid change for females, though there is a small proportion of cases even for GHQ score zero. The question is whether the GHQ score could be used to indicate the need for psychiatric treatment, and in particular

(A) What is the probability that a patient with GHQ=2 is a case?

(B) What is the GHQ score at which 70% of (all) patients are expected to be cases?

A fit of a generalized linear model with binomial errors and logit link function to these data gave a deviance of 4.942 (14 d.o.f.) when both the factor SEX and the covariate GHQ were included, and a deviance of 5.744 (15 d.o.f.) with only the covariate GHQ. The estimated coefficients in the latter model were $\hat{\alpha} = -3.454$ and $\hat{\beta} = 1.440$, and their estimated variance covariance matrix was

$$\begin{matrix} 0.317297 & -0.131591 \\ -0.131591 & 0.088771 \end{matrix}$$

(a) Explain why the second model can be accepted for the moment (in the absence of suitable residual plots). [4]

(b) Answer question (A) above, giving a 95% confidence interval. [9]

(c) Answer question (B) above, giving an approximate 95% confidence interval. [12]

HINT: If $X_0 = (\theta_0 - \alpha)/\beta$, where θ_0 is given, then the variance of $\hat{X}_0$ is approximately

$$\frac{1}{\beta^2}\left[\mathrm{Var}[\hat{\alpha}] - 2X_0\mathrm{Cov}[\hat{\alpha},\hat{\beta}] + X_0^2\mathrm{Var}[\hat{\beta}]\right].$$

Question 7 A random sample $y_1, \ldots, y_n$ is taken from a Gamma distribution $Ga(k, \theta)$ where $k > 0$ is known and $\theta > 0$ is the unknown parameter of interest.

(a) Find and sketch the likelihood function $L(\theta)$, up to a proportionality constant in θ. [4]

(b) Hence find the maximum likelihood estimate (MLE) $\hat{\theta}$ and show that the observed information at the MLE is $n\bar{y}^2/k$. [8]

(c) A Bayesian analysing these data has a prior for θ which is also Gamma, but with prior mean α/β and prior variance α/β^2. Show that the posterior distribution for θ is also Gamma, and write down the posterior mean and variance. [6]

(d) Show that the posterior mean is a weighted combination of prior mean and MLE, with the weight of the latter tending to 1 as n tends to infinity. [3]

(e) Comment on the posterior inference under great prior uncertainty (GPU), that is as α and β both tend to zero. [4]