UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B. Sc. Examination 2002

STATISTICS

ST52008A (ST219) Linear Models

Duration: 2 hours 15 minutes

Date and time:

Answer QUESTION 1 and any other TWO questions.

There are 75 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.

THIS EXAMINATION PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Question 1 You must answer this question. An experiment was held to determine whether or not four different tips produce different readings on a hardness testing machine. The machine operates by pressing the tip into a metal specimen. The hardness of the specimen can be determined by measuring the depth of the resulting depression. Four different specimens are used and the experimenter uses a randomised block design with specimen as the blocking factor. The results are given in the following table.

	Specimen			
Type of tip	1	2	3	4
1	9.3	9.4	9.6	10.0
2	9.4	9.3	9.8	9.9
3	9.2	9.4	9.5	9.7
4	9.7	9.6	10.0	10.2

Two students analysed these data using MINITAB. The session window and some graphs for each student are given below.

(a) Student 1 has correctly used a model for a randomised block design. Write down this model explaining what each term in the model represents and what assumptions were made. [5]

(b) Comment on the graphs student 1 has produced. What assumptions are being tested? Is there any reason to doubt the assumptions? [5]

(c) What conclusions should student 1 draw? [3]

(d) Explain how to find the fitted value for the first specimen and first tip. [3]

(e) Student 2 has analysed an incorrect model. What model is he using? What source of variation is he ignoring? [4]

(f) What conclusions would student 2 draw? [2]

(g) Student 2 has used Tukey's method to examine any differences between the tips. Explain why this method adjusts the significance level and how to interpret the output. [3]

```
STUDENT 1

Two-way ANOVA: y versus Specimen, Tip

Analysis of Variance for y
Source        DF        SS        MS        F        P
Specimen       3   0.82500   0.27500    30.94    0.000
Tip            3   0.38500   0.12833    14.44    0.001
Error          9   0.08000   0.00889
Total         15   1.29000


Plot RESI1 * FITS1

Macro is running ... please wait

Normal Prob Plot: RESI1


STUDENT 2

One-way ANOVA: y versus Tip

Analysis of Variance for y
Source     DF        SS        MS        F        P
Tip         3    0.3850    0.1283     1.70    0.220
Error      12    0.9050    0.0754
Total      15    1.2900
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  -----+---------+---------+---------+-
1           4     9.575     0.310      (---------*---------)
2           4     9.600     0.294       (---------*---------)
3           4     9.450     0.208  (---------*---------)
4           4     9.875     0.275                (---------*---------)
                                   -----+---------+---------+---------+-
Pooled StDev =    0.275               9.30      9.60      9.90     10.20
```

```
Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0117

Critical value = 4.20


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

                1           2           3

       2     -0.6017
              0.5517

       3     -0.4517     -0.4267
              0.7017      0.7267

       4     -0.8767     -0.8517     -1.0017
              0.2767      0.3017      0.1517


Plot RESI2 * FITS2

Macro is running ... please wait

Normal Prob Plot: RESI2
```

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]