UNIVERSITY OF LONDON

GOLDSMITHS' COLLEGE

B. Sc. Examination 2003

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. In the 1840's and 1850's the Scottish physicist James Forbes was interested in developing a method for estimating altitude on a hillside from measurement of the boiling point of water there. The temperature at which water boils is affected by atmospheric pressure, which in turn is affected by altitude. As part of this study Forbes collected the following data on atmospheric pressure (y measured in inches of mercury) and the boiling point of water (x measured in degrees farenheit) at 17 locations in Scotland and in the Alps.

y	x	y	x
20.79	194.5	20.79	194.3
22.40	197.9	22.67	198.4
23,15	199.4	23.35	199.9
23.89	200.9	23.99	201.1
24.02	201.4	24.01	201.3
25.14	203.6	26.57	204.6
28.49	209.5	27.76	208.6
29.04	210.7	29.88	211.9
30.06	212.2		

A statistician fitted a series of models to these data. Some extracts from the output are given below.

(a) Model 1 is a simple linear regression model. Write down this model and any necessary assumptions. [2]

(b) On the basis of the output it was decided to try omitting observation 12 to give Model 2. Explain why this was done and how it was achieved in MINITAB. [4]

(c) On the basis of the output it was decided to include a quadratic term in x. Explain why this was done. [2]

(d) What hypothesis is the Analysis of Variance Table for Model 3 testing? [2]

(e) Justify Model 3 as a better model than Model 2. [3]

(f) In Model 4 y is transformed to $\log_e y$ (ly). Why is this transformation considered? [2]

(g) Would you use Model 3 or Model 4 for prediction? Justify your answer. [3]

(h) Predict the value of y for another location with $x = 200.2$ using model 4. Explain how you would calculate a 95% prediction interval. (You should give the necessary formula but do not evaluate it.) [4]

(i) Define Cook's distance. How do we use it to determine if any observation has very high influence? [3]

Model 1

The regression equation is
y = - 81.1 + 0.523 x

Predictor	Coef	SE Coef	T	P
Constant	-81.064	2.052	-39.51	0.000
x	0.52289	0.01011	51.74	0.000

S = 0.2328 R-Sq = 99.4% R-Sq(adj) = 99.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	145.12	145.12	2677.11	0.000
Residual Error	15	0.81	0.05		
Total	16	145.94			

Unusual Observations

Obs	x	y	Fit	SE Fit	Residual	St Resid
12	205	26.5700	25.9201	0.0589	0.6499	2.89R

R denotes an observation with a large standardized residual

Model 2

Weighted analysis using weights in w

The regression equation is
y = - 80.7 + 0.521 x

16 cases used 1 cases contain missing values
or had zero weight

Predictor	Coef	SE Coef	T	P
Constant	-80.667	1.420	-56.81	0.000
x	0.520738	0.006997	74.42	0.000

S = 0.1608 R-Sq = 99.7% R-Sq(adj) = 99.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	143.15	143.15	5538.21	0.000
Residual Error	14	0.36	0.03		
Total	15	143.51			

Model 3

Weighted analysis using weights in w

The regression equation is
y = 117 - 1.42 x + 0.00475 x2

16 cases used 1 cases contain missing values
or had zero weight

Predictor	Coef	SE Coef	T	P
Constant	116.59	25.08	4.65	0.000
x	-1.4165	0.2462	-5.75	0.000
x2	0.0047522	0.0006040	7.87	0.000

S = 0.06951 R-Sq = 100.0% R-Sq(adj) = 99.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	143.449	71.724	14846.27	0.000
Residual Error	13	0.063	0.005		
Total	15	143.511			

Source	DF	Seq SS
x	1	143.150
x2	1	0.299

Unusual Observations

Obs	x	y	Fit	SE Fit	Residual	St Resid
14	209	27.7600	27.9059	0.0255	-0.1459	-2.26R

R denotes an observation with a large standardized residual

```
Model 4
Weighted analysis using weights in w

The regression equation is
ly = - 0.952 + 0.0205 x

16 cases used 1 cases contain missing values
                       or had zero weight

Predictor        Coef     SE Coef          T        P
Constant     -0.95177     0.02310     -41.20    0.000
x           0.0205186   0.0001138     180.24    0.000

S = 0.002616    R-Sq = 100.0%    R-Sq(adj) = 100.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1     0.22225     0.22225  32485.39    0.000
Residual Error    14     0.00010     0.00001
Total             15     0.22235
```

Question 2 A study in the 1970's in the US considered the variables which were important in determining the selling price of a house. The following independent variables were considered:

taxes	Property taxes (hundreds of dollars)
baths	Number of baths
area	Area of plot (thousands of square feet)
space	Living space (thousands of square feet)
garages	Number of garages
rooms	Number of rooms
bedrooms	Number of bedrooms
age	Age of home (years)
firepl	Number of fireplaces

The dependent variable (price) was the selling price in thousands of dollars. Data were collected on 28 houses.

(a) A statistician entered the data into MINITAB and produced the following output.

```
Stepwise Regression: sales versus taxes, baths, ...

Forward selection.  Alpha-to-Enter: 0.05

 Response is  sales   on  9 predictors, with N =   28


    Step          1        2        3
Constant    -1.4669  -0.6315  -1.0658

baths          31.3     17.8      8.2
T-Value       12.42     4.05     1.71
P-Value       0.000    0.000    0.100

taxes                   2.26     1.90
T-Value                 3.50     3.37
P-Value                0.002    0.003

space                            10.1
T-Value                          3.22
P-Value                         0.004

S              5.48     4.58     3.90
```

```
R-Sq          85.58    90.32    93.24
R-Sq(adj)     85.02    89.54    92.39
C-p            21.6      8.6      1.4
```

(i) Explain what fitting method is being used. [2]

(ii) Write down the model selected by this method. [1]

(iii) Comment on the fact that in the final model the coefficient for baths has a P-value of 0.100. [3]

(iv) Define the quantity R-Sq given in the output. Comment on its use to aid model choice. [3]

(v) Define the quantity R-Sq(adj) given in the output. In what way is it an improvement over R-Sq? [3]

(vi) Define the quantity C-p given in the output. Explain how it is used to aid model choice. [4]

(b) Explain the method of backwards fitting. What is a possible problem in fitting the full model with many variables? [6]

(c) Could the method of backwards fitting result in the same model as selected in part (a)? Justify your answer. [3]

Question 3 In an experiment to compare three different varieties of corn 15 plots in a field were allocated one of the three varieties at random. The yields (in bushels per plot) of the three varieties were as follows

Corn variety		
X	Y	Z
22.6	27.6	25.4
21.5	26.5	24.5
22.1	27.0	26.3
21.8	27.2	24.8
22.4	26.8	25.1

(a) Write down a suitable model for these data and any necessary assumptions. [4]

(b) Draw up the analysis of variance table and test for differences between the varieties. [7]

(c) Describe how to check the model assumptions using MINITAB

(i) graphically, [2]

(ii) using appropriate tests. [2]

(d) Y and Z are known to be related varieties and X is unrelated to the other two. Suggest appropriate mutually orthogonal contrasts for testing, explaining what hypothesis each of your contrasts tests, and carry out these tests. [10]

Question 4 A marketing organisation wishes to study the effects of four sales methods on weekly sales of a product. The organisation uses a randomised block design in which three salesmen use each of the sales methods. The results obtained are given below.

		Salesman		
		A	B	C
Method	1	32	29	30
	2	32	30	28
	3	28	25	23
	4	25	24	23

(a) Write down the model used and the assumptions made. [4]

(b) What is the advantage of using a randomised block design here rather than a completely randomised design? [4]

(c) Carry out a graphical analysis that shows that little interaction exists between sales method and salesmen. [6]

(d) Draw up the analysis of variance table and test for differences between the (i) sales methods and (ii) the salesmen. [11]