EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2011

MODULE 4 : Modelling experimental data

Time allowed: Three Hours

Candidates should answer **FIVE** *questions.*

All questions carry equal marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

The notation log *denotes logarithm to base* ***e****.*
Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Note also that $\binom{n}{r}$ *is the same as* ${}^nC_r$.

This examination paper consists of 11 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. (i) A randomised blocks experiment with b blocks is to be conducted on a standard treatment S and v other treatments A, B, … . Each block is to contain $(v + c)$ plots, consisting of c replicates of S and one of each of A, B, … .

(a) Write down a suitable mathematical model as the basis for analysing data from this experiment, stating any assumptions to be made about the terms in it.

(4)

[NOTE that for unequal replication the terms τ_i for treatment effects in the model satisfy the constraint $\Sigma r_i \tau_i = 0$, where treatment i is replicated r_i times.]

(b) Derive the least squares estimator of the difference between the effect of treatment S and that of any other treatment, and write down its variance.

(6)

(ii) A randomised blocks experiment was designed to compare four new (and possibly improved) methods of controlling a crop pest against the standard treatment. Each block contained 6 plots, consisting of 2 replicates of the standard treatment S and one of each new treatment A, B, C, D. The percentage damage, y, to the crop from each plot is given below.

		Treatment					*Total*
		S	A	B	C	D	
Block	I	22, 30	14	16	8	10	100
	II	24, 26	11	16	5	6	88
	III	18, 14	9	12	8	6	67
	IV	16, 12	8	6	4	5	51
Total		162	42	50	25	27	306

$\Sigma y^2 = 5076$

The researcher is only interested in comparing each new treatment with the standard treatment S.

(a) Construct an analysis of variance for these data. Calculate 95% confidence intervals for the difference in mean percentage crop damage between S and each new treatment. Comment on which of the new methods, if any, appear to result in a real improvement over the standard method.

(8)

(b) Discuss briefly any concerns there might be about carrying out the analysis in part (ii)(a).

(2)

2. There are 4 shelves in a fruit storage room, and it is possible to place 8 boxes of fruit side by side on each shelf. Four different storage times A – D are used in an experiment to measure the weight loss, y, of each box of fruit during storage. The experimental layout and the weight losses y (grams) are shown in the following table.

		Position								*Total*
		1	2	3	4	5	6	7	8	
Shelf	1	B 183	A 213	C 224	B 211	D 218	C 241	A 198	D 266	1754
	2	A 165	D 236	B 205	A 197	C 216	D 293	B 241	C 284	1837
	3	D 200	C 222	A 188	D 234	B 211	B 240	C 223	A 202	1720
	4	C 213	B 197	D 242	C 217	A 151	A 180	D 259	B 223	1682
Total		761	868	859	859	796	954	921	975	6993

$\Sigma y^2 = 1\,557\,997$ $\qquad \dfrac{6993^2}{32} = 1\,528\,189.031$

The treatment totals are: A, 1494; B, 1711; C, 1840; D, 1948.

(i) Explain carefully the steps in choosing a properly randomised design for this experiment.

(4)

(ii) Obtain the analysis of variance for these data.

(10)

(iii) Carrying out any statistical tests necessary to do so, write a report on the results of the experiment.

(6)

Turn over

3. An experiment was carried out to measure the effects of two factors on the maximum output voltage of a battery. Factor T was the operating temperature during manufacture, either 10 °C or 16 °C or 22 °C, and factor M was the material used in making the plates, either A or B. The experiment was carried out in completely randomised order with 3 replicates of each of the six treatment combinations, and voltages *y* were measured to the nearest 5 volts.

The following table gives the totals of the three replicates for each treatment combination.

Temperature	10°C	16°C	22°C	*Total*
Plate A	430	385	370	1185
Plate B	485	545	270	1300
Total	915	930	640	2485

The sum of the squares of all 18 observations was $\Sigma y^2 = 359\,825$.

The outline of the analysis of variance for these data is as follows, where the 3-level factor T is split into linear and quadratic components. The coefficients for constructing these components are: L (–1, 0, 1) and Q (1, –2, 1).

Source of variation	Degrees of freedom	Sum of Squares
Plates, M	1	***
Temperature, Linear T_L	1	***
Quadratic T_Q	1	***
Temperatures, T	2	8886.111
$M \times T_L$	1	***
$M \times T_Q$	1	***
Interaction M × T	2	5702.777
Total Treatments	5	15323.611
Residual	12	***
Total	17	***

(i) Construct a table showing the coefficients needed to calculate each of the five single-degree-of-freedom contrasts in the above analysis.

(4)

(ii) Complete the analysis of variance.

(9)

(iii) Draw a graph showing the totals (or the means) of all six treatment combinations.

(4)

(iv) Referring to the graph and the analysis of variance, recommend which treatment combinations should be used in a follow-up experiment.

(3)

4. An experiment is being designed in which five 2-level factors A, B, C, D, E are to be used. The scientist in charge would like to use a block size of 4, but is prepared to use blocks of size 8 if necessary. He feels that the smaller block size will lead to a more precise experiment but he wants to be able to estimate all the main effects and two-factor interactions. One complete replicate of the 32 treatment combinations {(1), *a*, *b*, … , *abcde*} is possible.

(i) Write down a suitable scheme using blocks of size 8, specifying

- the interactions confounded
- the principal block
- the other three blocks.

Give also the outline analysis of variance for a single replicate of this scheme.

(9)

(ii) Write down an optimal scheme using blocks of size 4, specifying

- the interactions confounded
- the principal block.

Give also the outline analysis of variance for a single replicate of this scheme.

(6)

(iii) Explain how a "residual" term may be constructed in each scheme, so that statistical tests can be made. What assumptions are required to do this?

(2)

(iv) Supposing that this experiment is one of a series using some or all of these factors, suggest any items of information that the scientist may have gathered earlier which would help to design the present experiment as efficiently as possible and meet all his requests.

(3)

5. In an experiment to investigate the absorption over time of rubidium and bromide ions into potato slices, five potato slices were immersed in a standard solution of rubidium bromide. The potato slices were each randomly assigned to one of five immersion durations. The table below gives the duration times of immersion in hours (Duration) and the corresponding observed uptakes of rubidium and bromide ions (in standardised units) of the potato slice (Absorption).

Duration (x)	$x - \overline{x}$	*Absorption*	
		Rubidium	*Bromide*
21.7	–42.38	7.2	0.7
46.0	–18.08	11.4	6.4
67.0	2.92	14.2	9.9
90.2	26.12	19.1	12.8
95.5	31.42	20.0	15.8

(i) Write down a single linear model for the absorption of both types of ions, using dummy variables and the centred duration times, $x - \overline{x}$. This model should allow for the fact that the intercept and slope of the relationship might be different for the two types of ions.

(4)

(ii) Obtain estimates of the parameters in your linear model, and show that the estimated error variance is 0.57. Test whether there is evidence against equality of the two slopes or intercepts.

(16)

6. The table below shows the results of a study of 21 children, investigating the link between age (in months) at first speaking a word and the score on an aptitude test.

Child (i)	*Age at first word (x_i)*	*Test score (y_i)*
1	15	95
2	26	71
3	10	83
4	9	91
5	15	102
6	20	87
7	18	93
8	11	100
9	8	104
10	20	94
11	7	113
12	9	96
13	10	83
14	11	84
15	11	102
16	10	100
17	12	105
18	42	57
19	17	121
20	11	86
21	10	100

$$\Sigma x_i = 302, \quad \Sigma y_i = 1967, \quad \Sigma x_i^2 = 5606, \quad \Sigma y_i^2 = 188155, \quad \Sigma x_i y_i = 26864.$$

(i) Analyse the data using a linear regression model. Include an analysis of variance table. What does your analysis indicate about the relationship between age at first word and test score? Construct a 95% confidence interval for the slope of the regression. (10)

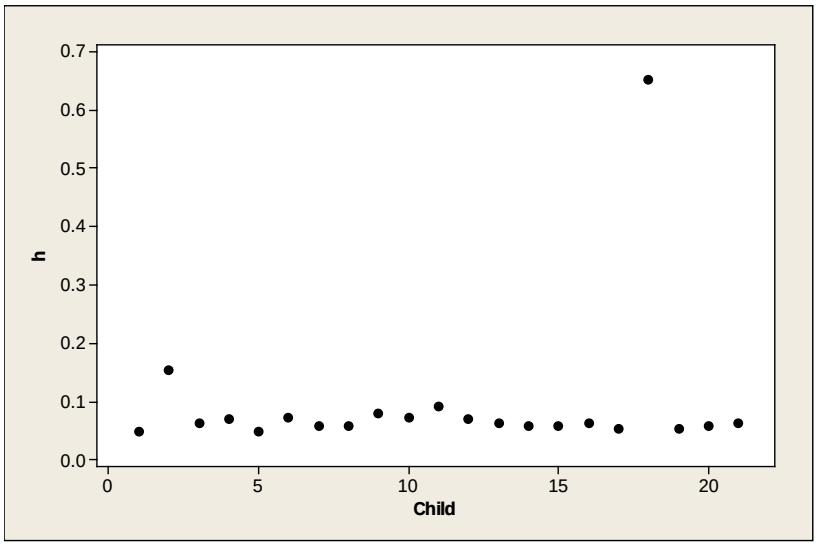
(ii) Define the *hat matrix*. What is an *influential observation*? How might the hat matrix be useful for identifying influential observations? (3)

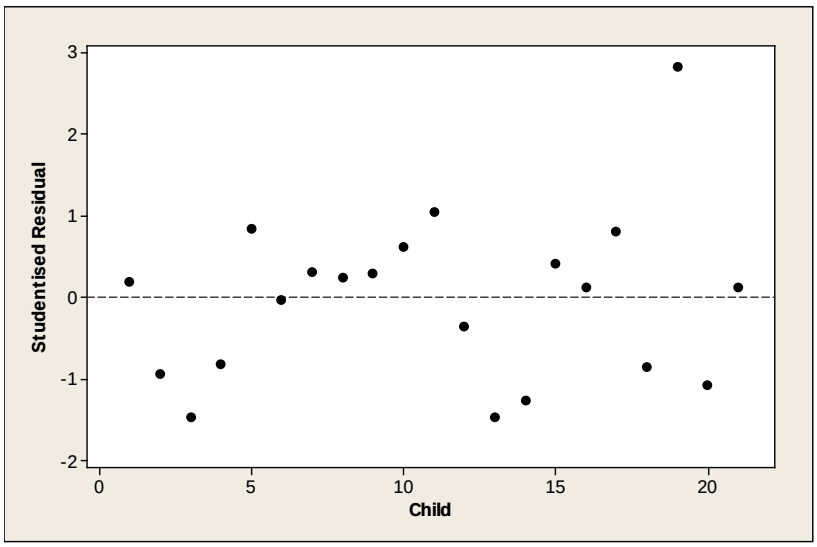
(iii) Define *Cook's distance*. (2)

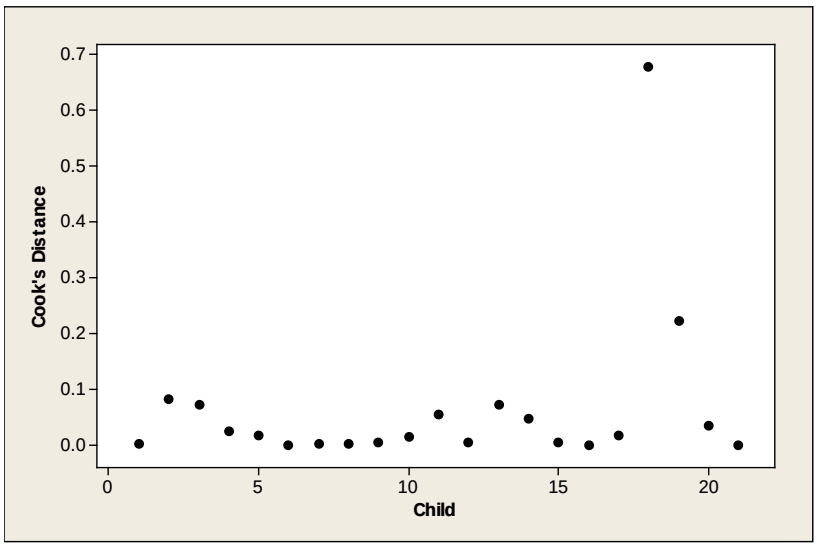
(iv) The graphs **on the next page** show plots of three regression diagnostics calculated from the data in the table: the diagonal elements, *h*, of the hat matrix; the studentised residuals; Cook's distance. Summarise the results of the diagnostics. How, if at all, do these diagnostics affect your conclusions in part (i)? (5)

Graphs for question 6 are on the next page

Graphs for question 6







Turn over

7. The average rent in $ per acre of land (Y) planted with alfalfa was thought possibly to depend on the following four regressors:

X_1 the average rent in $ per acre for all tillable land,

X_2 the density of dairy cows (number per square mile),

X_3 the proportion of farmland used for pasture,

X_4 which takes the value 1 if liming is required to grow alfalfa, 0 otherwise.

Data were collected to investigate this possibility. The sample size was 67. For regression analysis, Y was transformed to the log scale.

(i) Computer output of part of the analysis of these data is shown below. Use this output to decide whether any of the regressors can be excluded from the regression model.

(5)

```
             Estimate     Std. Error
(Intercept)  2.115843      0.125314
X1           0.024885      0.001850
X2           0.014175      0.002894
X3           0.018319      0.318821
X4           0.161466      0.076370

Residual standard error: 0.2496
R-squared: 0.8486,  Adjusted R-squared: 0.8389
```

(ii) The table below shows the residual sum of squares (RSS) for models including each possible subset of the regressors. The total sum of squares is 25.5147. Use the technique of stepwise regression, with a 5% significance level, to find the most parsimonious model for these data.

(11)

Model	*RSS*	*Model*	*RSS*
X_1	6.0660	X_2, X_4	21.6286
X_2	22.9544	X_3, X_4	22.8577
X_3	23.3396	X_1, X_2, X_3	4.1406
X_4	25.5150	X_1, X_2, X_4	3.8623
X_1, X_2	4.1426	X_1, X_3, X_4	5.3562
X_1, X_3	5.3567	X_2, X_3, X_4	15.1378
X_1, X_4	5.9121	X_1, X_2, X_3, X_4	3.8621
X_2, X_3	15.6070		

(iii) Why, in general, might the approaches to model selection in parts (i) and (ii) lead to different conclusions?

(4)

8. (i) Define the *deviance* of a generalised linear model. How can deviances be used to compare two models, M_1 and M_2, when M_2 is nested within M_1?

(3)

(ii) The data below show the number of people in certain administrative districts of Queensland, Australia, who were strip-searched by the police in a recent year, and the number of them who were not subsequently charged with any offence.

District	*Sex*	*Strip-searched*	*Not charged*
Gympie	M	172	100
Gympie	F	13	6
Oxley	M	302	166
Oxley	F	46	30
Rockhampton	M	2057	1266
Rockhampton	F	219	111
Townsville	M	91	57
Townsville	F	20	15
Warwick	M	127	93
Warwick	F	18	14

The computer output **on the next page** shows an analysis of these data using a binary logistic regression model for the probability of not being charged following a strip-search. In the analysis of deviance table, the five districts have been entered as a single factor (District).

(a) What can you conclude about the effects of Sex and District on the probability of not being charged following a strip-search?

(14)

(b) Construct a 95% confidence interval for the odds in favour of not being charged following a strip-search for a male in Oxley.

(3)

Output for question 8 is on the next page

Output for question 8

Coefficients:

Predictor	Estimate	Std. Error	z-value	P(>|z|)
Constant	0.07385	0.18607	0.40	0.69143
District				
Oxley	−0.02497	0.18404	−0.14	0.89205
Rockhampton	0.13877	0.15480	0.90	0.37002
Townsville	0.34611	0.24884	1.39	0.16426
Warwick	0.75523	0.24058	3.14	0.00169
Sex				
M	0.23700	0.12050	1.97	0.04921

Null deviance: 28.8485 on 9 degrees of freedom

Residual deviance: 9.5067 on 4 degrees of freedom

AIC: 70.522

Analysis of Deviance Table

Model: binomial, link: logit

Terms added sequentially (first to last)

	Df	Deviance
NULL	9	28.8485
District	5	13.3434
Sex	4	9.5067

Variance-Covariance Matrix

	Interc	Oxley	Rockham	Townsv	Warwick	Male
Intercept	0.0346	-0.0230	−0.0225	−0.0237	−0.0230	−0.0135
Oxley	−0.0230	0.0339	0.0221	0.0222	0.0222	0.0009
Rockhamp	−0.0225	0.0221	0.0240	0.0222	0.0221	0.0004
Townsv	−0.0237	0.0222	0.0222	0.0619	0.0222	0.0017
Warwick	−0.0230	0.0222	0.0221	0.0222	0.0579	0.0009
Male	−0.0135	0.0009	0.0004	0.0017	0.0009	0.0145