EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2012

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 $\genfrac{}{}{0pt}{}{n}{r}$ *is the same as* ${}^{n}C_{r}$.

This examination paper consists of 14 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. A researcher has collected data concerning the effect of different types of study (reading books or attending lectures) on a student's final mark in an examination. The response variable was the final mark (y) and the explanatory variables were number of books read (x_1) and number of lectures attended (x_2). 40 students, regarded as a random sample, participated in the investigation. The data are summarised below.

$$\sum_{i=1}^{40} y_i = 2542 \qquad \sum_{i=1}^{40} y_i^2 = 172\,428 \qquad \sum_{i=1}^{40} x_{1i} = 80 \qquad \sum_{i=1}^{40} x_{1i}^2 = 240 \qquad \sum_{i=1}^{40} x_{2i} = 564$$

$$\sum_{i=1}^{40} x_{2i}^2 = 8666 \qquad \sum_{i=1}^{40} y_i x_{1i} = 5543 \qquad \sum_{i=1}^{40} y_i x_{2i} = 37\,186 \qquad \sum_{i=1}^{40} x_{1i} x_{2i} = 1234$$

(i) Fit a simple linear regression model with number of books (x_1) as the explanatory variable. Construct an analysis of variance table for this model and test the hypothesis that the slope equals zero against a two-sided alternative.

(6)

(ii) Computer output from fitting a multiple regression model for final mark (MARK) with explanatory variables number of books (BOOKS) and number of lectures attended (ATTEND) is shown below. Find the four values that have been deleted and replaced by (1) – (4).

```
Regression Analysis: MARK versus BOOKS, ATTEND

The regression equation is
MARK = 37.4 + 4.04 BOOKS + 1.28 ATTEND

Predictor     Coef  SE Coef      T      P
Constant    37.379    7.745   4.83  0.000
BOOKS        4.037    1.753    (1)  0.027
ATTEND      1.2835   0.5870   2.19  0.035

S = 14.0522   R-Sq = (2)%   R-Sq(adj) = (3)%

Analysis of Variance

Source          DF       SS      MS     F      P
Regression       2   3577.7  1788.8   (4)  0.001
Residual Error  37   7306.2   197.5
Total           39  10883.9
```

(4)

(iii) Construct the analysis of variance table that has sources of variation "BOOKS alone", "ATTEND adjusted for BOOKS" and "Residual". What can be concluded about the effect of BOOKS and ATTEND on MARK from this table?

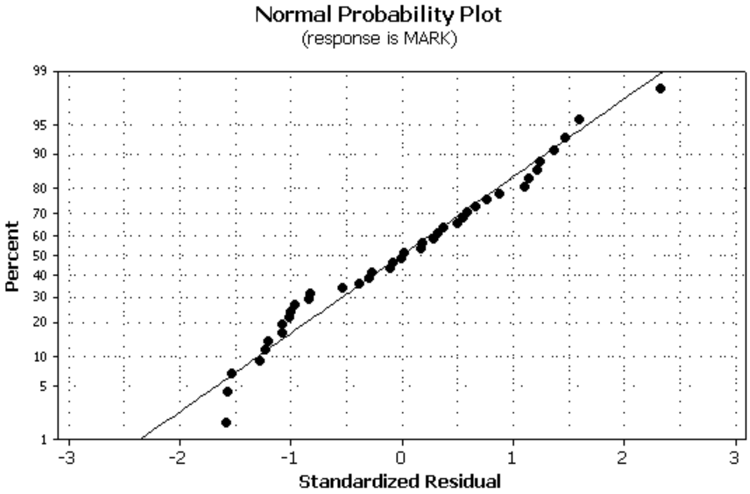
(7)

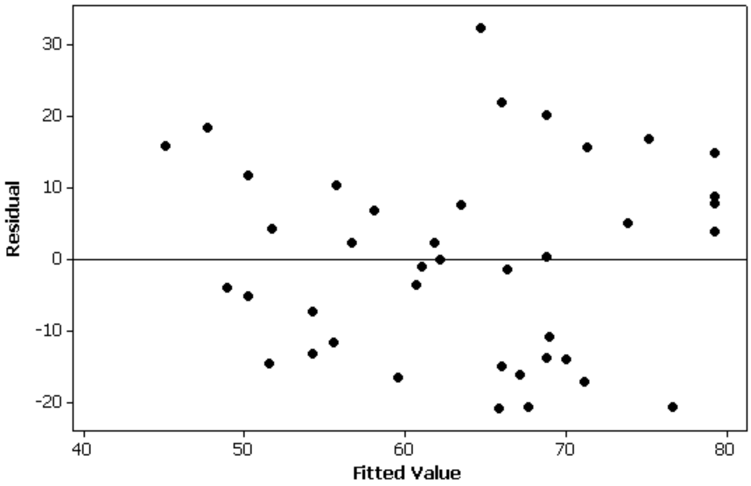
Question continued on next page

Turn over

(iv) Some additional output from the analysis is given below. Give the formula for calculating a standardised residual. What can you conclude from the graphs about the suitability of the model in part (ii) for these data?

(3)





2. The yields of a chemical from two industrial processes were measured at 10 different temperatures: 65, 66, 67, 68, 69, 70, 71, 72, 73 and 74 degrees, these values being denoted below by $x_1, x_2, \ldots, x_{10}$.

Letting $y_{1,i}$ (i = 1, 2, ..., 10) denote the yield of Process 1 at temperature x_i and $y_{2,i}$ (i = 1, 2, ..., 10) denote the yield of Process 2 at temperature x_i, the results obtained are summarised as follows.

$$\sum_{i=1}^{10} x_i = 695 \qquad \sum_{i=1}^{10} x_i^2 = 48\,385 \qquad \sum_{i=1}^{10} y_{1,i} = 836.5 \qquad \sum_{i=1}^{10} y_{1,i}^2 = 70\,065.73$$

$$\sum_{i=1}^{10} x_i y_{1,i} = 58\,220.6 \qquad \sum_{i=1}^{10} y_{2,i} = 783.6 \qquad \sum_{i=1}^{10} y_{2,i}^2 = 61\,515.06 \qquad \sum_{i=1}^{10} x_i y_{2,i} = 54\,551.9$$

(i) A model involving two parallel regression lines

$$y_{g,i} = \alpha_g + \beta(x_i - \bar{x}) + \varepsilon_{g,i} \qquad g = 1, 2; \quad i = 1, 2, \ldots, 10$$

is considered appropriate for these data, where $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$ and $\varepsilon_{1,1}$, $\varepsilon_{1,2}$, …, $\varepsilon_{1,10}$, $\varepsilon_{2,1}$, $\varepsilon_{2,2}$, …, $\varepsilon_{2,10}$ are mutually independent Normally distributed random variables with mean 0 and variance σ^2.

Write down the forms of $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ if this model is expressed as a general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Derive expressions for estimates of the parameters α_1, α_2, β and σ^2 using standard results for the general linear model. Hence calculate estimates of these parameters for the given data.

(10)

(ii) The model

$$y_{g,i} = \mu + \delta z_g + \beta(x_i - \bar{x}) + \varepsilon_{g,i} \qquad g = 1, 2; \quad i = 1, 2, \ldots, 10$$

was fitted to the same data, where $\varepsilon_{g,i}$ has the same distribution as before. The variable z_g is called Process in the output and is coded as 1 for Process 1 and 2 for Process 2. Selected output for this analysis is presented below. Show how the five values printed in boldface may be obtained from your calculations in part (i).

(5)

```
Variable       Estimate   Std Error
Intercept      88.94      5.60
Temperature     1.06      0.0799
Process        -5.29      0.459

S = 1.026
```

Residual Sum of Squares (RSS)

Model	*Yield = Temperature + Process*	*Yield = Temperature*
RSS	17.894	157.815

Question continued on next page

Turn over

(iii) Suppose now that a model involving the same regression line for both processes is proposed for the relationship between temperature and yield. Making use of figures from your calculations or from the computer output, or from both, test the hypothesis that such a model is appropriate, using a *t* test and an *F* test. What do your results illustrate about the relationship between a *t* test and an *F* test in a simple linear regression model?

(5)

3. A study was conducted to investigate models for the monthly man-hours needed to maintain the anaesthesiology service for each of twelve naval hospitals in the USA. Three possible explanatory variables X1 (the number of surgical cases), X2 (the eligible population per thousand) and X3 (the number of operating rooms) were recorded for each hospital.

(i) The correlation matrix of the explanatory variables is given below.

```
              X1        X2        X3
X1         1.000     0.960     0.926
X2         0.960     1.000     0.940
X3         0.926     0.940     1.000
```

The eigenvalues of the correlation matrix are 2.884, 0.077 and 0.038. What does this tell us about multicollinearity in the data?

(3)

(ii) Briefly describe how variable selection can be used to address the problem of multicollinearity.

(3)

(iii) The table below shows the Residual Sums of Squares (RSS) for all possible models. Use the table to perform a stepwise regression starting from the model including no variables.

(9)

Residual Sum of Squares

Included variables	*RSS*
None	15 094 263
X1	590 734
X2	748 192
X3	1 094 190
X1, X2	371 975
X1, X3	269 077
X2, X3	465 202
X1, X2, X3	219 125

Question continued on next page

Turn over

(iv) The table below shows parameter estimates for all possible models. Using this table and your previous results, write a brief summary of your findings about the relationship between man-hours and the explanatory variables.

(5)

Regression coefficient estimates and standard errors for all models

	Intercept		*X1*		*X2*		*X3*	
Model	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
None	1582.35	338.16						
X1	142.21	115.63	4.21	0.27				
X2	180.66	128.38			9.43	0.68		
X3	–520.83	209.02					315.48	27.89
X1, X2	130.59	96.85	2.43	0.81	4.17	1.81		
X1, X3	–176.80	127.38	2.67	0.51			127.02	38.72
X2, X3	–153.20	178.17			5.78	1.66	131.38	56.15
X1, X2, X3	–114.59	130.34	2.03	0.68	2.27	1.68	99.73	42.22

4. The following data were collected after a food poisoning outbreak. It is suspected that the potato salad, the crab salad or both were the cause. The contingency table below shows the results of a random survey of 304 diners: whether they were sick (food-poisoned) and the food that they ate.

	Potato Salad		*No Potato Salad*	
	Crab Salad	*No Crab Salad*	*Crab Salad*	*No Crab Salad*
Not Sick	80	24	31	23
Sick	120	22	4	0

(i) What is a *generalised linear model*? What is the *saturated model* in the context of generalised linear models?

(4)

(ii) A log-linear generalised linear model with a Poisson distribution was fitted to the data. The computer output below shows the analysis of deviance table for these data with parameter estimates for several models. Each row of the table refers to a model containing the terms given in the left-hand column of that row and all the rows above it.

```
                       Deviance      Change in Deviance
intercept               295.253
sick                    294.779         0.474
potato                  169.664       125.115
crab                     73.871        95.793
potato:crab              63.196        10.676
sick:potato               6.482        56.714
sick:crab                 2.743         3.739
sick:potato:crab      4.123e-10         2.743
```

Find a suitable model for these data and give an interpretation. What can be concluded about the likely cause of the outbreak?

(6)

(iii) How is a *Pearson residual* defined in this model?

(2)

(iv) Calculate the Pearson residuals for your fitted model. Do they indicate an adequate fit to the data?

(4)

(v) A colleague suggests that a logistic regression model with sickness as response would be more appropriate for these data than the log-linear model. Describe briefly the different aims of these two approaches, and discuss whether your colleague's suggestion is a good one.

(4)

5. (i) Explain the differences between the assumptions made when a *random effects* model is used instead of a *fixed effects* model when analysing a set of data. Discuss briefly how the results of an analysis of variance calculation are used in each case.

(6)

(ii) An investigation has been carried out to measure the percentage of fibre in Soya Cotton Cake, with special interest in studying the extent of variation during routine chemical analysis of samples of the same product. Three laboratories took part in the investigation and three different technicians carried out the analysis at each laboratory. Both the laboratories and the technicians were randomly selected from larger numbers available. Each technician carried out four analyses. The data were coded for ease of analysis, and a summary of the results is given in the following table.

Totals of 4 results obtained by Technicians

Laboratory A			*Laboratory B*			*Laboratory C*		
Tech A1	*Tech A2*	*Tech A3*	*Tech B1*	*Tech B2*	*Tech B3*	*Tech C1*	*Tech C2*	*Tech C3*
2.30	1.50	3.64	4.15	3.41	3.10	3.35	3.59	3.24
Laboratory total: 7.44			Laboratory total: 10.66			Laboratory total: 10.18		

The total of all 36 analyses was 28.28 and the total sum of squares of the 36 analyses was 23.8436.

(a) Construct an analysis of variance from which the components of variance at each stage (laboratories and technicians) can be estimated, and calculate these estimates.

(10)

(b) Use your results to comment on whether any changes to the sampling scheme in part (ii), either in numbers of technicians taking part or in number of analyses carried out by each technician, could usefully be made when planning further investigations of this type.

(4)

6. (i) Explain fully the circumstances in which a *balanced incomplete block* design is appropriate.

(3)

(ii) Derive the two equations that must be satisfied for a balanced incomplete block design to exist, having the parameters v, b, r, k and λ (in the usual notation). State what each parameter represents.

(3)

(iii) Six treatments A–F are to be compared for their effect on the flavour of a food, and members from a trained tasting panel will undertake the work. The flavourings are such that not more than 5 samples can be tasted by a panel member at each session. With the facilities available, a maximum of N = 30 samples can be examined at each session; each panel member is a "block". Three possible schemes for planning a session to compare the 6 treatments are to use block sizes (a) 2, (b) 3, (c) 5.

Write down an experimental design for each of (a), (b), (c), stating the values of the parameters listed in part (ii).

(3, 4, 2)

(iv) The residual variance σ^2 for scheme (b) is thought to be 15% higher than for scheme (a), and that for scheme (c) is thought to be 40% higher than for scheme (a). Which design would you recommend? Give reasons for your answer.

(5)

[The variance of the difference between two treatment means in a balanced incomplete block design is $\dfrac{2k\sigma^2}{\lambda v}$.]

Turn over

7. An experiment has been carried out to investigate six factors A, B, C, D, E, F, each at two levels (low and high). A quarter-replicate of a 2^6 design was used, with E equated to ABC and F equated to BCD. Data, in suitably coded units, were as follows.

Treatment combination	(1)	ae	bef	abf	cef	acf	bc	abce
Response	8.1	11.0	9.9	15.1	11.2	11.5	10.0	9.9

Treatment combination	df	adef	bde	abd	cde	acd	bcdf	abcdef
Response	17.3	15.6	13.5	9.2	9.7	14.8	12.4	16.4

It was known from earlier work that E and F did not interact with one another, nor with any of the other factors. It was also known that C and D did not interact with one another.

(i) Write down the defining contrast for this scheme, and use it to list all the alias sets.

(6)

(ii) Using the information about absence of interactions, indicate which main effects and interactions can be estimated from these alias sets. Stating clearly what assumptions must be made, identify any alias sets which can provide degrees of freedom for residual in an analysis of variance.

(4)

(iii) The sum of the squares of the 16 observations listed above is 2513.72, and the sum of the observations is 195.6. A partial analysis of variance table is as follows.

Item	*D.F.*	*Sum of squares*	*F ratio*
A		8.1225	
B		0.4900	
C		0.9025	
D		30.8025	
E		0.0900	
F			
AB		0.2025	
AC		3.2400	
AD		1.6900	
BC		2.1025	
BD		5.0625	
Residual			
Total			

Copy and complete this table.

(4)

Question continued on next page

Turn over

(iv) Show how the sum of squares for the AB interaction would have been computed (you may assume the arithmetic is correct).

(1)

(v) Determine which, if any, of the main effects and interactions are significant.

(2)

(vi) One of the components (with 1 d.f.) included in "residual" accounts for a sum of squares equal to 33.0625. Examine the result of removing this component from "residual" and discuss

(a) the validity of this removal,

(b) the significance of main effects and interactions after removal of this component.

(3)

8. An experiment has been carried out into the growth of a corn crop in a region. Two small sites were available near to a research centre. Each site could accommodate 3 replicates of a randomised complete block design that contained 5 treatments. The same treatments were used at each site, and the staff at each site carried out the experiment. The treatments were a standard, locally used, method of cultivation of the crop (S), 3 different variations on this (W, X, Y) and a more mechanised method (M) which had not previously been used in the region. Site 1 undertook some training in the use of M but Site 2 did not.

Crop records at the end of the experiment are given (in suitable units) in the following table.

Treatment	*S*	*M*	*W*	*X*	*Y*	*Block total*
Site 1, block I	18.6	21.3	23.5	18.6	15.4	97.4
II	23.2	28.4	22.8	17.7	25.9	118.0
III	21.0	32.1	28.2	19.2	20.0	120.5
Site 2, block IV	21.3	16.0	23.0	16.0	18.8	95.1
V	15.5	12.5	27.6	20.2	20.9	96.7
VI	14.8	21.3	18.4	14.0	19.1	87.6

Treatment totals on the two sites are as follows.

	S	*M*	*W*	*X*	*Y*	*Site total*
Site 1	62.8	81.8	74.5	55.5	61.3	335.9
Site 2	51.6	49.8	69.0	50.2	58.8	279.4
Overall total	114.4	131.6	143.5	105.7	120.1	$G = 615.3$

The sum of the squares of all 30 observations is 13 242.39, and the corrected sum of squares for blocks I–VI is 180.131.

The scientist in charge of the experiments asks for an analysis whose outline is shown in the following table.

Source of variation	*Degrees of freedom*	*Sum of squares*
Sites		
Blocks within sites		
Treatments		145.975
Residual		
TOTAL		

(i) Write down, and explain carefully, the linear model he is using. State the properties of each term in the model, and the assumptions for the model to be valid.

(3)

(ii) Copy and complete the analysis of variance table.

(7)

Question continued on next page

Turn over

(iii) Calculate the treatment means at each site, and show these on a graph using different symbols for each site. Comment on any points this raises about the analysis above.

(4)

(iv) Discuss how the linear model given in part (i) might usefully have been modified for this experiment when it was being planned. Recalculate the analysis to match any new model you propose.

(4)

(v) Briefly list the points you would make when writing a report on this experiment. (A set of "bullet points" will suffice.)

(2)