EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2009

(Modular format)

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 12 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. In many instances of linear modelling, a response variable y might be dependent on more than one predictor variable. Thus a set of variables x_i (i = 1, 2, …, p) could be used to predict y through the general linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

where the β_i are model parameters and ε is an error term.

(i) State clearly all the assumptions made in fitting such a model.

(2)

(ii) Write down the equivalent matrix formulation of the model, and state the form of the least squares estimators for the parameters in the model.

(3)

(iii) These least squares estimators have some very useful properties.

(a) State the properties they possess irrespective of the distribution of the errors.

(b) State the extra properties they possess if the errors are independent and Normally distributed.

(3)

(iv) Explain why highly dependent predictor variables can cause problems in fitting such a model. What methods can be used to try to overcome such problems?

(3)

(v) Explain why an adjusted R^2 value is often preferred to R^2 when comparing models.

(2)

(vi) Explain what is meant by *influential observations* and why they can be a problem. Describe some of the diagnostics that can be obtained from statistical packages to detect influential observations.

(7)

2. (i) Explain fully the circumstances in which a *balanced incomplete block* design is appropriate. In the usual notation let v represent the number of treatments, b the number of blocks, r the number of replicates of each treatment and k the number of units in each block. Also let λ denote the number of times each pair of treatments occurs together in a block.

Derive two equations connecting the parameters v, b, r, k, λ. (5)

(ii) In an industrial experiment, 5 batches of metal ingots were selected at random from the output of a production process; each batch contained 4 ingots. Different amounts of cadmium (Cd) and tin (Sn) were used in five treatments, as follows.

A	10% Cd	No Sn
B	20% Cd	No Sn
C	30% Cd	No Sn
D	10% Cd	10% Sn
E	30% Cd	10% Sn

Each ingot was melted and mixed with the appropriate amount of Cd and Sn and then allowed to cool. When the treated ingot was reheated, its melting point y (^{o}C) was recorded, as shown below.

Batch 1	A:	194	D:	205	E:	250	B:	214
Batch 2	B:	204	E:	243	D:	198	C:	238
Batch 3	D:	206	B:	205	C:	238	A:	186
Batch 4	A:	183	E:	247	B:	202	C:	229
Batch 5	E:	255	C:	244	D:	209	A:	198

$\Sigma y = 4348$, $\Sigma y^2 = 955360$.

In parts (b) and (c) you may use the following results.

- The treatments sum of squares adjusted for which batches they appear in is $\Sigma Q_i^2/vk\lambda$, and $Q_A = -421$, $Q_B = -142$, $Q_C = 311$, $Q_D = -215$, $Q_E = 467$
- The adjusted treatment means are $\bar{y}_i + Q_i/\lambda v$
- The variance of a difference between any two adjusted means is $2k\sigma^2/v\lambda$.

(a) State the values of v, b, r, k and λ in the above design. (2)

(b) Obtain the analysis of variance for the data. State briefly your conclusions from this. (7)

(c) Estimate the adjusted treatment means, compare these means and interpret your results. (6)

3. An experiment was conducted in pots in a controlled environment to examine the effects of five soil treatments A – E, each at 2 levels, on the growth of barley plants. Factor A was volcanic ash ("high" level) or mineral acid soil; B was acidity, high or low pH; C was potassium present ("high" level) or absent; D was magnesium present ("high" level) or absent; E was calcium oxide present ("high" level) or absent.

A fractional factorial design was used, consisting of 8 treatment combinations as shown in the table below (in the usual notation for 2-level experiments). One pot per treatment combination was used, and the mean height (cm) of the plants growing in each pot after a fixed period of time is shown in the table.

Treatment Combination	*e*	*ad*	*bde*	*ab*	*cd*	*ace*	*bc*	*abcde*
Growth (cm)	8.7	12.0	17.5	11.0	9.0	13.0	16.1	17.7

(i) Verify that the defining contrast consisted of I = ACE and I = BDE.

(4)

(ii) Write down the complete defining contrast and the full set of aliases for this design.

(4)

(iii) Construct a table showing the coefficients required to estimate each of the main effects A, B, C, D, E and the interactions AB, AD. Hence estimate the main effects and interactions, so far as is possible in this design.

(5)

(iv) The experimenter comments that two of these effects seem larger than the others, but cannot interpret the two-way table of means for AB. Explain how this table is calculated, and why it is not helpful to attempt an interpretation of it.

(3)

(v) It is sometimes argued that effects which appear small should be pooled to produce a residual sum of squares with more degrees of freedom. Comment on the advisability of doing this in fractional factorial designs.

(4)

4. (i) Distinguish between *fixed* and *random* effects in modelling experimental data. Explain briefly when each should be used.

(4)

(ii) A company is studying the variability in tensile strength of the steel beams that it produces. The beams are produced at three different sites, and some of the variability may be due to differences between sites. Another source of variability could be differences between batches of steel used to produce the beams.

The company has selected four batches at random at each site, and from the production of each batch three beams have been selected at random and their tensile strengths have been measured.

(a) A model suggested for these data is

$$y_{ijk} = \mu + s_i + b_{ij} + \varepsilon_{ijk} .$$

Interpret each of the terms in this model and state clearly the assumptions needed to conduct an analysis.

(5)

(b) A partially completed analysis of variance is as follows.

Source of variation	Sum of squares	Expected mean square
Between sites	418.72	$\sigma^2 + 3\sigma_b^2 + 6\Sigma s_i^2$
Between batches within sites	1701.59	$\sigma^2 + 3\sigma_b^2$
Within batches	250.00	σ^2

Complete this analysis and report on the importance of the two possible sources of variability. Your report should contain details of how any necessary estimates have been made and of any hypotheses that have been tested.

(11)

5. An investigation on the fuel consumption of cars was carried out, to examine the effects of the size of engine and the speed at which a car was driven, and the interaction between them. Similar cars were used in all trial runs, and all other conditions were kept as far as possible the same in each run. Readings were obtained on fuel consumption, y miles per gallon, for three runs at each combination of engine size and speed of driving, as shown in the table below.

	Size of engine (cc)				
Speed (mph)	1100	1500	1800	Σy	Σy^2
30	43.8, 45.2, 44.9	38.0, 38.6, 37.4	37.1, 35.5, 35.2	355.7	14185.91
50	43.0, 42.1, 43.4	41.7, 41.6, 39.9	41.0, 40.0, 41.6	374.3	15577.99
70	32.6, 31.4, 32.4	29.9, 28.0, 29.0	29.9, 31.1, 31.8	276.1	8489.95
Σy	358.8	324.1	323.2	1006.1	
Σy^2	14580.94	11913.19	11759.72		38253.85

(i) Write down a suitable model for explaining these data, giving the meanings of all the terms. State any assumptions necessary to carry out an analysis.

(4)

(ii) Draw a graph of the means of engine/speed combinations.

(3)

(iii) Complete an analysis of variance for these data.

(8)

(iv) Using (ii) and (iii), write a report, suitable for a non-statistician, on the effects of the two factors, engine size and driving speed, on fuel consumption.

(5)

Turn over

6. An economist is studying salaries for employees in a large company. He has data on 434 employees. The variables are as follows.

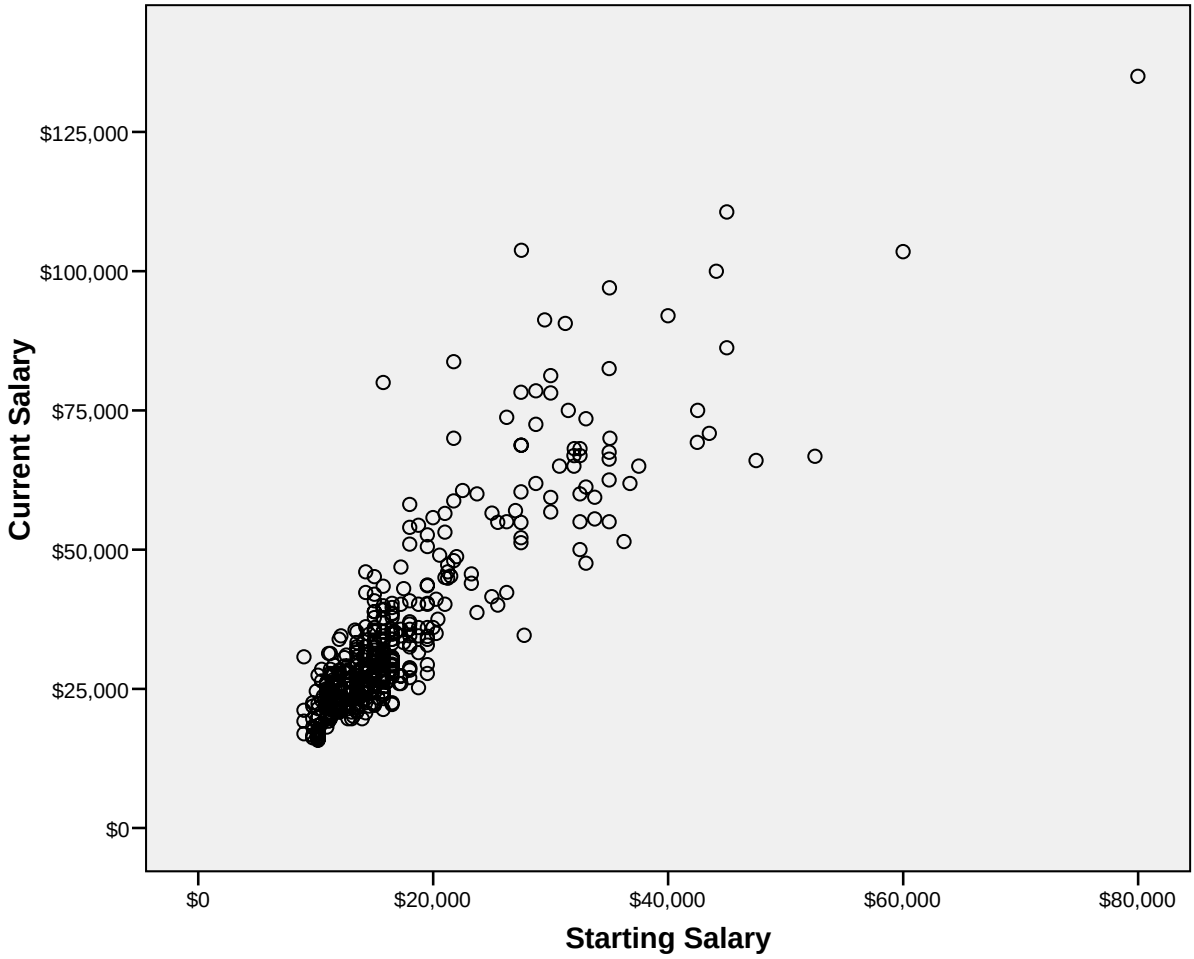
Current salary in dollars

Starting salary, when the employee joined the company, in dollars

Sex of employee

Grade of employee (clerk or team leader or manager)

The economist wants to construct a model to predict current salary from the other variables, and is convinced that current salary is related to starting salary. He shows the following scatter plot as evidence of this relation.



Question 6 is continued on the next two pages

Turn over

(i) Someone tells the economist that it would be better if he scaled all the salary values by dividing by 1,000 to make the numbers more manageable. He is worried that this might affect the analysis, changing the values and statistical significance of the parameter estimates obtained. What would your advice be and why?

(2)

(ii) Someone else advises the economist to model the logarithm of the current salary value rather than the raw value. Why might this be a good idea?

(2)

(iii) The economist suggests that if he is using the logarithm of the current salary then maybe he should use the logarithm of the starting salary. What would you advise?

(3)

(iv) The economist has a computer program that will do multiple regression, but he does not know how to model factors using this program. Write down the form of a suitable multiple regression model to include all three predictor variables, and using the logarithms of the salary values. Explain each term in the model and how it relates to the original variables.

(4)

(v) The economist obtains a multiple regression model using the logarithms of both of the variables describing salaries. The response variable is called logsal. He produces a plot of the standardised residuals against the predicted values obtained from the model, and a plot of the standardised residuals against the logarithm of starting salary. Interpret these plots in a way that he could understand. **The plots are shown on the next page.**

(6)

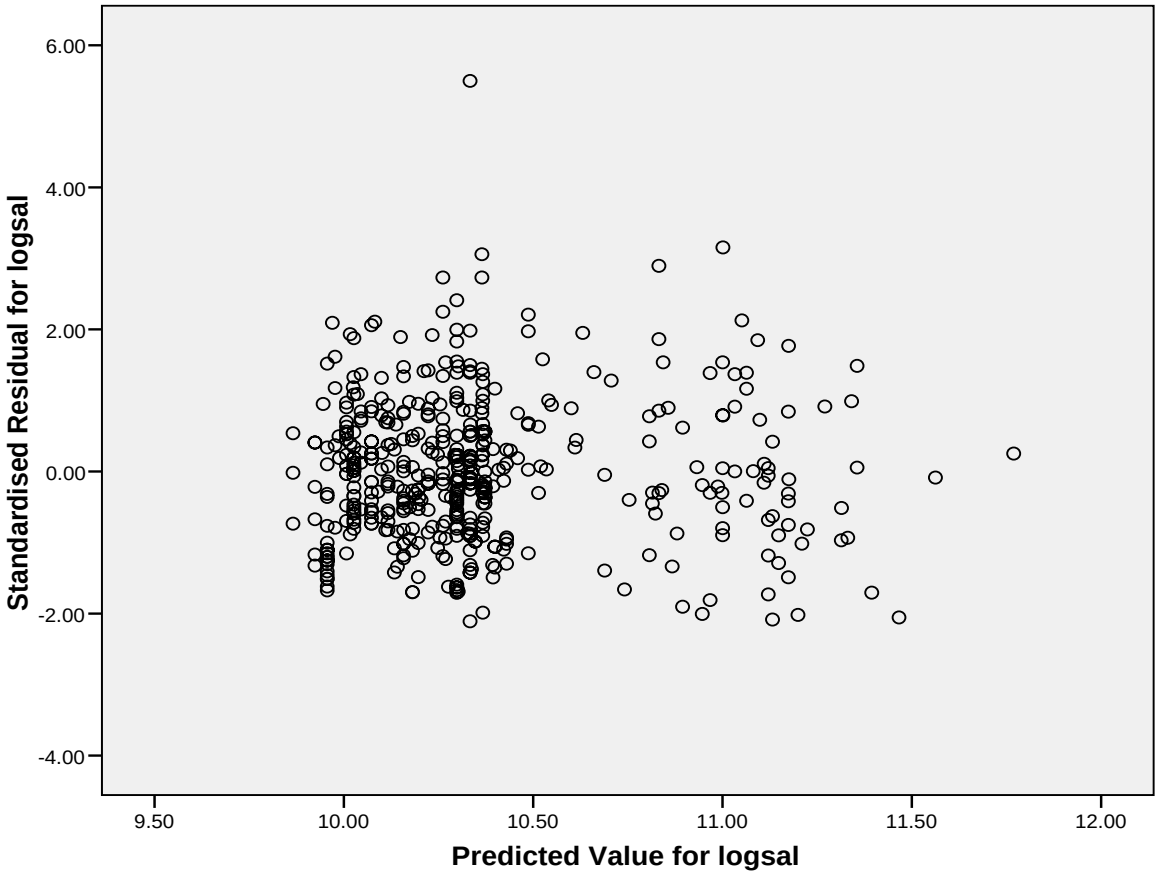
(vi) Someone else says that they have found an interaction between job category and logarithm of starting salary.

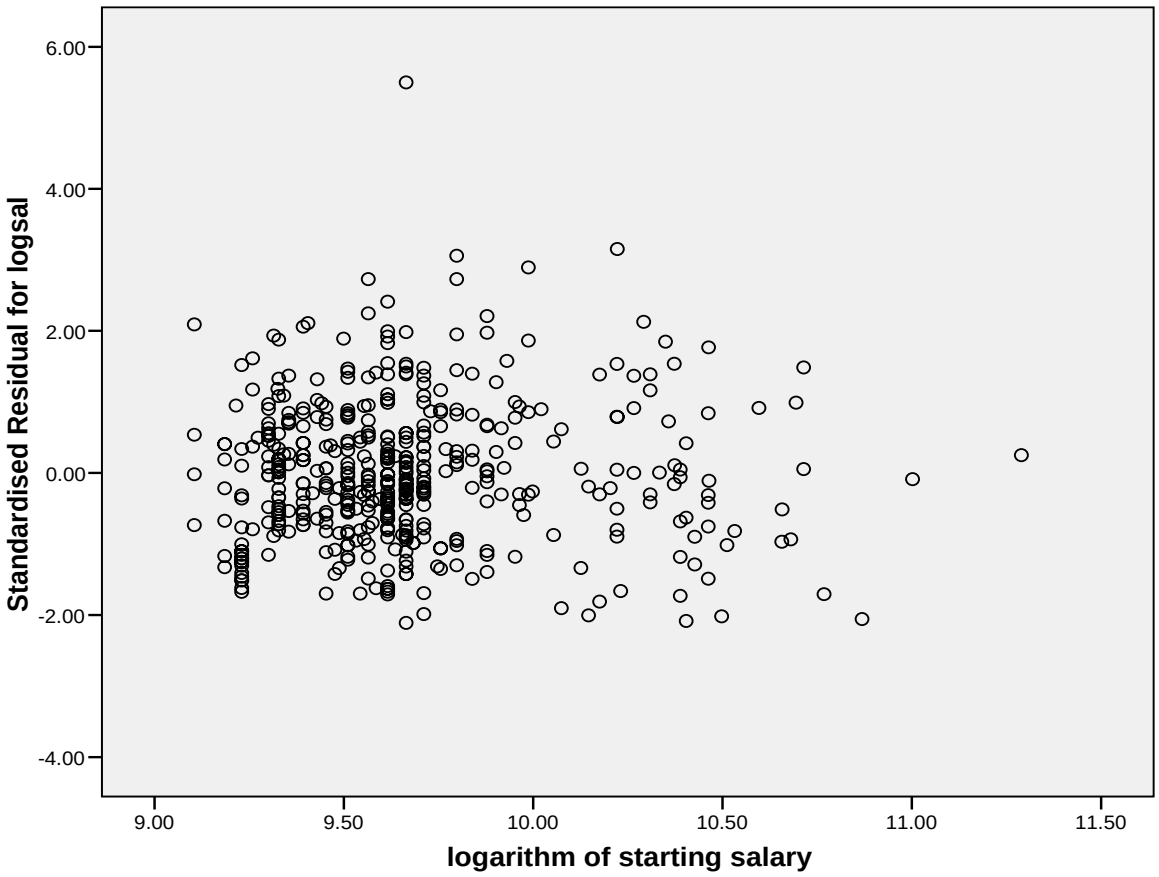
(a) Show how this could be included in your multiple regression model.

(b) Explain what such an interaction would mean.

(3)

Question 6 is continued on the next page, which shows the plots for part (v)


Standardised Residual for logsal
6.00
4.00
2.00
0.00
-2.00
-4.00
9.50
10.00
10.50
11.00
11.50
12.00
Predicted Value for logsal


Standardised Residual for logsal
6.00
4.00
2.00
0.00
-2.00
-4.00
9.00
9.50
10.00
10.50
11.00
11.50
logarithm of starting salary

7. (i) Briefly discuss the relative merits of forward selection and backward elimination as applied to model selection in multiple linear regression.

(5)

(ii) Three process variables $X1$, $X2$ and $X3$ can be adjusted in a chemical plant, and each variable might affect the yield Y from the plant. Twenty-one observations have been made on Y at different values of $X1$, $X2$ and $X3$. It is required to find the best combination of $X1$, $X2$ and $X3$ for prediction of the yield, Y. The table below shows residual sums of squares from various linear models fitted to the data.

Variables in model	Residual sum of squares
–	2069.24
$X1$	319.12
$X2$	483.15
$X3$	1738.44
$X1$, $X2$	188.80
$X1$, $X3$	309.14
$X2$, $X3$	475.06
$X1$, $X2$, $X3$	178.83

(a) Use forward selection to choose a model. Show your working.

(3)

(b) Use backward elimination to choose a model. Show your working.

(3)

(c) Which model would you recommend and why?

(2)

(d) What other information would you like to have in order to suggest a "good" model?

(3)

(e) How would you check that your final chosen model was a good fit to the data?

(4)

Turn over

8. A doctor is investigating the effect of a woman's age on the success of an IVF (in vitro fertilisation) procedure. She has randomly selected 10 women aged under 35 and 10 women aged at least 35. From hospital records she has obtained the following data, which record the numbers of eggs obtained from the women and the numbers that were fertilised during one IVF procedure. She wants to investigate the effect of the woman's age on the probability of an egg being successfully fertilised. She calls this probability the "fertilisation rate".

Women aged under 35		**Women aged at least 35**	
Number of eggs	*Number of fertilised eggs*	*Number of eggs*	*Number of fertilised eggs*
10	9	7	6
9	7	10	7
7	5	9	5
5	3	8	4
10	9	6	4
7	7	5	1
9	5	7	4
8	8	6	4
7	2	5	2
7	5	7	5

(i) Carry out a suitable exploratory analysis to see whether the fertilisation rate might depend on the woman's age.

(4)

(ii) Let n_i denote the number of eggs and x_i the number of fertilised eggs for the ith woman. Let π_i denote the fertilisation rate for the ith woman.

(a) Explain why a binomial distribution may be valid to model the data.

(2)

(b) Write down the expression for the log likelihood of the observed data, assuming a binomial distribution with different fertilisation rates for each woman. Identify the logit function in your expression.

(2)

Question 8 is continued on the next page

(iii) The data are analysed using a generalised linear model, with the logit link. The model assumes constant fertilisation rate within each age group, so contains a constant and age as a covariate. Age is coded as 1 for older women, and 0 for younger women. Part of the output from a computer program is given below.

```
Deviance  =  28.26     (1/df) Scaled Deviance  =  1.57

Variance function:  V(u) = u*(1-u/eggs)      [Binomial]
Link function    :  g(u) = log(u/(eggs-u))    [Logit]
```

(a) Explain why the highlighted value 1.57 is useful, and how it is derived from the other numerical value in the output.

(2)

(b) Explain what the highlighted expressions V(u) and g(u) are and how their formulae are obtained.

(2)

(c) The estimated value of the coefficient for age in the generalised linear model is –0.744 and the estimate of the constant is 1.150.

Obtain estimates of the predicted success rates for the two types of women.

(4)

(d) For the model which contains only the constant (i.e. does not take age into account), the value for the scaled deviance is 32.65. State, with reasoning, whether the effect of woman's age is statistically significant.

(2)

(e) Someone else has modelled these data but coded younger women as age = 1 and older women as age = 0. Explain how the results and estimates would be different from those given above.

(2)