EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2013

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* ${}^{n}C_{r}$.

This examination paper consists of 16 printed pages.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. An experiment is to be conducted to compare the responses of nine treatments A – I. There is sufficient material available for 10 replicate samples to be processed using each of the nine treatments, and up to ten samples could be processed in a day. It is anticipated that there will be differences in the processing from day to day. An initial step is needed to prepare the material for processing, but while each run of this initial step can only produce enough material for a maximum of seven samples, batches of samples from this initial step can be stored for processing at a later date.

(i) Identify the advantages and disadvantages associated with using a randomised complete block design or an incomplete block design for this experiment.

(5)

(ii) Explain the following relationships required for a balanced incomplete block design.

$$rt = bk$$

$$\lambda(t-1) = r(k-1)?$$

Here, t is the number of treatments, r is the number of replicates, b is the number of blocks, k is the number of units per block and λ is the number of times that each pair of treatments appears together in a block.

(2)

(iii) One of the possible balanced incomplete block designs that could be constructed to cope with the constraints described above for this experiment has the following parameters as defined in part (ii): $t = 9$, $r = 10$, $k = 5$, $b = 18$, $\lambda = 5$. Find the parameter values for three other possible such designs. Discuss which of these four designs would make most effective use of the available resources.

(6)

(iv) Describe how you would construct the allocation of treatments to blocks for the balanced incomplete block design with parameters $t = 9$, $r = 10$, $k = 5$, $b = 18$, $\lambda = 5$, and write down the allocation of treatments to blocks for this balanced incomplete block design.

(7)

2. (i) Briefly discuss the differences between a confounded factorial design and a replicated fractional factorial design, highlighting the advantages and disadvantages of each approach.

(5)

A factorial experiment is to be performed using six factors, A – F, each at two levels. However, resources only allow for 64 experimental units to be included in the experiment, with a further constraint that only 8 experimental units can be processed during each day, so that the experiment will take 8 days to complete.

(ii) Identify an appropriate confounding scheme to allow estimation of all main effects and two-factor interactions in a confounded factorial design for this experiment, clearly showing the contents of the principal block. From this write down also the contents of the other blocks.

(6)

(iii) Write down the outline of the analysis of variance table, listing the terms in it and their degrees of freedom. What assumption would you need to make to assess the significance of the main effects and two-factor interactions?

(4)

Before starting the experiment it is realised that practical constraints associated with the application of the treatments mean that two of the factors, E and F, need to be kept constant during each day (changing the levels of these factors doubles the amount of time taken to process each experimental unit).

(iv) Briefly describe how this would influence the design of the confounded factorial experiment, and the structure of the analysis of variance table, commenting on the information that could be obtained about the main effects of and interaction between these two factors, and the impact on the estimation of the other main effects and two- and three-factor interactions.

(5)

3. An experiment to assess the impact of plant density (plants per square metre) on the production of a flower crop, measured as the mean plant dry weight, was arranged following a Latin square design in a large glasshouse compartment. The glasshouse compartment was built with an external wall facing due South, a door from a corridor on the North side, and with adjoining compartments to the East and West. From previous experiments it is expected that there might be variability in both temperature and light levels along both the North-South and East-West axes of the compartment.

(i) Explain how a Latin square design allows the elimination of any systematic variation along the North-South and East-West axes of the compartment, and write down the linear model that is the basis for analysing data from this experiment, stating the properties of each term in the model.

(4)

The arrangement of treatments and mean plant dry weights for each experimental unit are shown **on the next page**, the first number in each cell of the table being the density in plants per square metre and the second number the mean plant dry weight in grams. Also given is the analysis of variance table and a table of means.

(ii) Comment on the overall Density term and on the importance, or otherwise, of including the North-South (row) and East-West (column) blocking terms in the analysis model.

(3)

(iii) Construct a 95% confidence interval for the mean plant dry weight at a density of 10 plants per square metre, and a 95% confidence interval for the difference in mean plant dry weights between densities of 20 and 30 plants per square metre.

(4)

The first two orthogonal polynomial contrasts, representing linear and quadratic trends, have been included in the analysis to allow assessment of the shape of the response with increasing density.

(iv) Define a linear contrast between the totals $\{T_i\}$ for v treatments, each of which is replicated r times, and state the sum of squares (with one degree of freedom) associated with this contrast in an analysis of variance.

(3)

(v) Explain the benefits of including mutually orthogonal contrasts if this is possible, and write down the conditions for two contrasts, L and Q, to be mutually orthogonal, in terms of the coefficients l_i and q_i.

(3)

(vi) Interpret the extra information provided by the inclusion of the two orthogonal polynomial contrasts.

(3)

Data for Question 3 are on the next page

	NORTH					
WEST	25 259.2	30 228.3	20 278.5	10 360.3	15 310.3	EAST
	20 287.8	10 375.2	25 250.9	15 345.1	30 292.9	
	15 355.6	25 297.8	30 265.0	20 338.1	10 408.2	
	30 348.4	15 364.4	10 413.6	25 295.5	20 331.7	
	10 434.6	20 375.9	15 383.5	30 319.4	25 354.5	
	SOUTH					

Analysis of variance table

Variate: mean plant dry weight

Source of variation	df	SS	MS	MS ratio	*p*-value
row stratum	4	22694.1	5673.5	16.09	
col stratum	4	1394.9	348.7	0.99	
row.col stratum					
Density	4	41079.0	10269.8	29.12	<.001
Lin	1	37911.6	37911.6	107.51	<.001
Quad	1	2905.7	2905.7	8.24	0.014
Deviations	2	261.7	130.9	0.37	0.698
Residual	12	4231.5	352.6		
Total	24	69399.6			

Tables of means

Variate: mean plant dry weight

Grand mean 331.0

Density	10	15	20	25	30
	398.38	351.78	322.40	291.58	290.80

Standard errors of differences of means

Table	Density
rep.	5
d.f.	12
s.e.d.	11.876

4. In a field experiment to assess the efficacy of a number of new pesticides, data were collected in an experiment, arranged as a randomised complete block design, on the numbers of aphids found on a set of sampled plants both before the application of the pesticides and 2 days after application. In addition to six different new pesticide treatments (two of which consist of combinations of two of the three individual pesticides – i.e. treatment AC consisted of both of pesticides A and C, and BC consisted of both of pesticides B and C), an industry standard pesticide and an untreated control treatment were included (with the untreated control, in which no sprays were applied, having two replicate plots in each block). The data are shown below.

Treatment	*Block*	*Before*	*After*	*Treatment*	*Block*	*Before*	*After*
A	1	25	10	BC	1	91	6
A	2	39	40	BC	2	34	9
A	3	70	15	BC	3	34	3
A	4	55	8	BC	4	33	2
B	1	31	16	Standard	1	18	19
B	2	162	6	Standard	2	33	25
B	3	70	12	Standard	3	49	32
B	4	31	4	Standard	4	36	10
C	1	14	8	Untreated	1	69	38
C	2	37	16	Untreated	2	30	28
C	3	24	21	Untreated	3	52	4
C	4	45	43	Untreated	4	12	25
AC	1	63	15	Untreated	1	85	56
AC	2	53	23	Untreated	2	37	34
AC	3	114	7	Untreated	3	70	68
AC	4	39	6	Untreated	4	6	22

The output **on the next page** shows the result of analysing the counts of aphids two days after application of the pesticides using a generalised linear model (GLM) which enables overdispersion to be detected if it is present.

Question continued on the next page

Accumulated analysis of deviance

Change	df	deviance	mean deviance	deviance ratio	*p*-value
+ block	3	13.891	4.630	0.56	0.644
+ treatment	6	173.647	28.941	3.53	0.013
Residual	22	180.585	8.208		
Total	31	368.122	11.875		

Dispersion parameter is estimated to be 8.21 from the residual deviance.

Parameters for factors are differences compared with the reference level for each factor:

Factor	Reference level
block	1
treatment	Untreated

Estimates of parameters

Parameter	estimate	s.e.	*t* value	*p*-value	antilog of estimate
Constant	3.600	0.256	14.05	<.001	36.61
block 2	0.075	0.307	0.24	0.810	1.077
block 3	-0.036	0.315	-0.12	0.909	0.9643
block 4	-0.336	0.342	-0.98	0.336	0.7143
treatment A	-0.633	0.377	-1.68	0.107	0.5309
treatment B	-1.286	0.496	-2.59	0.017	0.2764
treatment C	-0.446	0.351	-1.27	0.217	0.6400
treatment AC	-0.992	0.437	-2.27	0.033	0.3709
treatment BC	-1.928	0.664	-2.91	0.008	0.1455
treatment Standard	-0.469	0.354	-1.33	0.199	0.6255

(The s.e. values are based on the residual deviance.)

(i) Identify the components of a generalised linear model, and describe the particular form of generalised linear model that is appropriate for these data.

(5)

(ii) Interpret the results of the analysis. In particular, identify those pesticides that significantly reduced the level of aphid infestation, and estimate the percentage reduction provided by each pesticide treatment.

(7)

(iii) How might you modify the analysis to take account of the impacts of the combined applications of pairs of pesticides? Are there any changes to the design that you would recommend to the experimenter with regard to these combined treatments?

(5)

(iv) Describe what is meant by *overdispersion*, and indicate how evidence for overdispersion could be identified in the analysis output.

(3)

5. An experiment was performed to investigate the effect of weed competition and phosphate fertiliser on the yield of maize for two different varieties (labelled A and B). There were two weed competition treatments (F = weed-free, W = weed-infested) and four fertiliser rates (O = control, P1 = 20 kg phosphate per ha, P2 = 40 kg phosphate per ha, P3 = 60 kg phosphate per ha). The experiment was arranged as a randomised complete block design in two blocks, each block containing 16 equal sized plots. Within each block, the 16 treatment combinations were allocated at random to plots.

The tables below summarise the total yields for the two plots for each treatment combination (in coded units), plus the totals for different combinations of factor levels.

Variety		A	A	B	B
Weed Competition		W	F	W	F
Fertiliser Rate	O	19.1	20.4	25.8	28.8
	P1	21.3	22.6	32.9	37.7
	P2	27.0	29.4	42.3	46.9
	P3	26.8	28.7	39.7	45.1
Totals		94.2	101.1	140.7	158.5

Totals		Variety		Weed Competition		
		A	B	W	F	Totals
Fertiliser Rate	O	39.5	54.6	44.9	49.2	94.1
	P1	43.9	70.6	54.2	60.3	114.5
	P2	56.4	89.2	69.3	76.3	145.6
	P3	55.5	84.8	66.5	73.8	140.3
Totals		195.3	299.2	234.9	259.6	494.5

The two block totals (for 16 plots each) are 256.8 and 237.7, and the sum of squares for the 32 observations is 8259.57. You may also use the facts that the sum of squares of the 16 treatment totals ($19.1^2 + 21.3^2 + \ldots + 45.1^2$) is 16 478.49, that the sum of squares for the eight totals for Variety/Fertiliser Rate combinations is 32 861.87, and that the sum of squares for the eight totals for Weed Competition/Fertiliser Rate combinations is 31 503.25.

Question continued on the next page

(i) Construct an analysis of variance to assess the effects of variety, weed competition and fertiliser treatment (and the interactions between these factors) on the yield of maize.

(7)

(ii) Extract the sum of squares for the linear single-degree-of-freedom contrast for the fertiliser factor [the coefficients of the linear contrast for four equally-spaced levels of a factor are (–3, –1, 1, 3)].

(2)

(iii) Draw a diagram showing the treatment means for all 16 combinations of variety, weed competition and fertiliser, illustrating how the effects of fertiliser treatments vary with variety and weed competition.

(4)

(iv) Using the diagram and the analysis of variance, explain the results found by this experiment, including the quantitative response to changes in levels of applied phosphate, and any important interactions between the three factors (variety, weed competition and fertiliser rate).

(7)

6. In a study of the effect of altitude changes on systolic blood pressure, data were collected on 38 male individuals born at high altitude in a primitive environment in the Peruvian Andes, who had then moved into the mainstream of Peruvian society living at a much lower altitude. Particular interest lay in the suggestion that migration from a primitive society to a modern one might result in increased blood pressure at first, with a subsequent decrease to normal levels. For each individual, the response variable is systolic blood pressure (*sbp*) with eight explanatory variables also collected – years since migration (*years*), *age* (in years), *weight* (in kg), *height* (in mm), chin skin fold (in mm – *chin*), forearm skin fold (in mm – *forearm*), calf skin fold (in mm – *calf*) and pulse rate (in beats per minute – *pulse*).

(i) Describe the idea of *parsimony* with regard to selecting models in a multiple linear regression analysis. (2)

(ii) Write down the least-squares estimator of the parameter vector $\boldsymbol{\beta}$ in the usual general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. State the Gauss-Markov theorem concerning this estimator. (3)

(iii) Write down the normal equations for a linear regression model with two explanatory variables, and solve these to give the parameter estimates for the coefficients associated with each of the explanatory variables. Identify how you would obtain the constant parameter having obtained these estimates. (3)

The output **on the next page** shows the results of using a forward selection stepwise procedure with the variance ratios for including or excluding variables from the model both set to have the value 3, and the model summary and parameter estimates for the selected model. Further output shows the model summary and parameter estimates for the best model when considering an all possible subset selection process, based on the Adjusted R-squared statistic.

(iv) Briefly describe how stepwise selection approaches work, considering both the *forward selection* and *backward elimination* methods, and identifying how choice of values for the variance ratios for the inclusion and exclusion of variables influences the final model selection. (4)

(v) Define the *Adjusted R-squared statistic*, commenting on how it can be used in the selection of the best model through comparison of all possible models. State other statistics that might be used to identify whether an additional term should be added to an existing model. (4)

(vi) Interpret the analysis output for the forward selection stepwise process, including an interpretation of the estimated parameters, commenting on how the best fitting model from the all possible subsets process differs from that identified using the stepwise process. (4)

Output for Question 6 is on the next page

Forward Selection Stepwise Process

Values are the residual mean squares as a result of making the indicated change to the current model, with the changes sorted by increasing value of the residual mean square.

```
Step 1:  95.92 – Adding weight   122.44 – Adding forearm   122.87 – Adding age
        125.23 – Adding calf     126.21 – No change        126.29 – Adding height
        127.84 – Adding chin     129.52 – Adding years     129.56 – Adding pulse
Chosen action: Adding weight

Step 2:  87.82 – Adding years     93.66 – Adding chin       94.27 – Adding pulse
         95.92 – No change        98.08 – Adding height     98.34 – Adding age
         98.43 – Adding forearm   98.66 – Adding calf      126.21 – Dropping weight
Chosen action: Adding years

Step 3:  84.31 – Adding chin      87.82 – No change         88.40 – Adding pulse
         88.60 – Adding height    89.38 – Adding forearm    89.83 – Adding age
         89.85 – Adding calf      95.92 – Dropping years   129.52 – Dropping weight
Chosen action: No change
```

Source	df	SS	MS	MS ratio
Regression	2	1596.0	798.07	9.09
Residual	35	3074.0	87.82	
Total	37	4670.0	126.21	

Percentage variance accounted for 30.4;
Standard error of observations estimated as 9.37.

Parameter	estimate	s.e.	*t* value
Constant	62.6	15.1	4.15
weight	1.104	0.260	4.25
years	-0.383	0.184	-2.08

Summary for Model with the highest Adjusted R-squared value

Change	df	SS	MS	MS ratio	*p*-value
+ weight	1	1216.80	1216.80	15.11	<.001
+ years	1	379.34	379.34	4.71	0.037
+ chin	1	207.26	207.26	2.57	0.118
+ height	1	202.06	202.06	2.51	0.123
+ pulse	1	87.72	87.72	1.09	0.304
Residual	32	2576.64	80.52		
Total	37	4669.82	126.21		

Percentage variance accounted for 36.2;
Standard error of observations estimated as 8.97.

Parameter	estimate	s.e.	*t* value	*p*-value
Constant	139.7	48.8	2.86	0.007
weight	1.694	0.350	4.84	<.001
years	-0.418	0.184	-2.27	0.030
chin	-1.443	0.698	-2.07	0.047
height	-0.0588	0.0343	-1.71	0.096
pulse	-0.181	0.174	-1.04	0.304

7. (i) Discuss the differences between linear and non-linear regression models, including a description of the different approaches that need to be taken to fit each type of model.

(4)

(ii) State whether each of the following models is linear or non-linear, identifying the linear and non-linear parameters in each case, and also identifying whether any non-linear models can be transformed into a linear form.

(4)

(a) $y = a + \dfrac{c}{1 + \exp(-b(x - m))}$

(b) $y = a + bx + cx^2 + dx^3$

(c) $y = \dfrac{ax}{1 + bx}$

(d) $y = ae^{bx}$

A field experiment was performed to assess the effect of adding phosphate fertiliser to the production of sugar in sugar beet. An exponential model (with parameters A, B and R, and with X representing the explanatory variable) was fitted to describe the observed response, and the results from this model fitting are shown **on the next page**, including a plot of the fitted model and observed data.

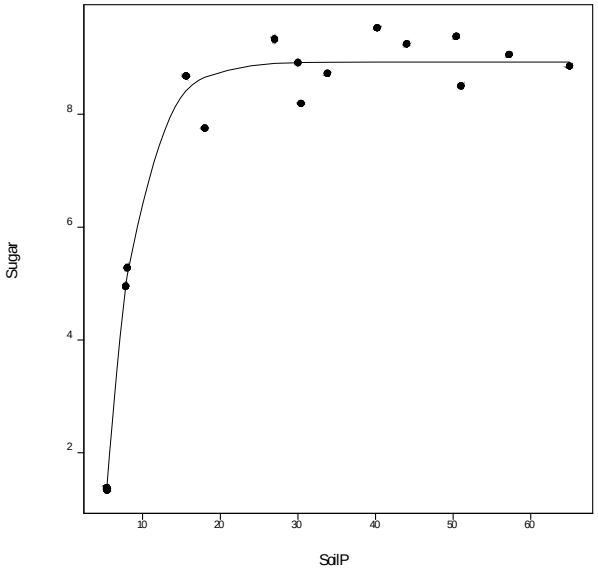
(iii) Interpret the results of the analysis, including a description of how the fitted parameters relate to the sugar response to phosphate. Estimate the level of phosphate needed to achieve a sugar level of 8.5.

(7)

(iv) Describe the concepts of *leverage* and *influence* in regression analysis for the usual general linear model $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, and state how leverage values are obtained from the "hat" matrix $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Describe why the four observations identified in the output have high leverage values.

(5)

Output for Question 7 is on the next page



Nonlinear regression analysis

```
Response variate:    Sugar
     Explanatory:    SoilP
    Fitted Curve:    A + B*(R**X)
     Constraints:    R < 1
```

Summary of analysis

Source	df	SS	MS	MS ratio
Regression	2	109.076	54.5380	279.56
Residual	13	2.536	0.1951	
Total	15	111.612	7.4408	

Percentage variance accounted for 97.4;
Standard error of observations is estimated to be 0.442.

The following units have large standardised residuals.

Unit	Response	Residual
5	7.757	-2.15

The following units have high leverage.

Unit	Response	Leverage
1	1.338	0.49
2	1.384	0.49
3	4.953	0.42
4	5.281	0.44

Estimates of parameters

Parameter	estimate	s.e.
R	0.7681	0.0266
B	-31.31	6.35
A	8.928	0.135

8. An experiment has been performed to obtain detailed information about the expression of a gene believed to be associated with the development of mushrooms. Samples were collected daily from three replicate growing trays for each of three different varieties of mushrooms (A, B and C), grown at three different temperatures (18, 21 and 24°C) over the 9 day period that the mushroom development process generally takes. The data are summarised in the following graphs – each row of graphs shows the data for a different variety (A in the top row, B in the middle row, C in the bottom row), and each column of graphs shows the data for a different temperature (18°C in the left column, 21°C in the centre column, 24°C in the right column). The gene expression response essentially counts the number of copies of each gene in each sample, expressed on a log (base 2) scale.

Linear regression analysis was used to assess whether the pattern of gene expression was linear with time, and to identify any impacts of the different treatment combinations on the parameters of the linear response. The presented output includes the accumulated analysis of variance obtained from fitting a sequence of models, and the fitted parameter estimates for two of the three models fitted as part of the analysis. In the model definitions, "gexp" is the vector of gene expression values, "time" is the vector of observation times (in days), and "treat" is the factor indicating the treatment combination associated with each observation; "~" is used to indicate that the response depends on the model specified on the right hand side, with "*" indicating the crossing of the two terms, i.e. that the model contains both main effects and the associated interaction.

(i) Describe the sequence of models that have been fitted to the data, identifying how consecutive models in the sequence differ.

(4)

(ii) Explain why model 3 is the most suitable to describe the observed responses, and interpret the fitted parameters for this model.

(5)

(iii) Given the set of treatments considered in this experiment, briefly describe how the analysis might be extended to provide more detailed information about the impacts of variety and temperature on gene expression.

(3)

(iv) Describe how the analysis could be extended to provide a test of the lack-of-fit of the model relative to the between-replicate variation.

(3)

(v) Identify the assumptions required for this analysis, and explain how you would calculate and analyse the residuals to examine these assumptions.

(5)

Output and graphs for Question 8 are on the next two pages

Accumulated Analysis of Variance Table

Model 1: gexp ~ 1 (null model)
Model 2: gexp ~ time
Model 3: gexp ~ time + treat
Model 4: gexp ~ time * treat

Model	Res.df	RSS	df	SS	MS ratio	*p*-value
1	269	1806.91				
2	268	880.52	1	926.39	301.6192	<0.001
3	260	807.19	8	73.33	2.9843	0.003
4	252	773.99	8	33.20	1.3510	0.219

Model 2: gexp ~ time

Coefficients:

	Estimate	Std. Error	*t* value	*p*-value
(Intercept)	5.51202	0.20503	26.88	<0.001
time	0.64490	0.03841	16.79	<0.001

Residual standard error: 1.813 on 268 degrees of freedom
Multiple R-squared: 0.5127, Adjusted R-squared: 0.5109
MS ratio: 282 on 1 and 268 df, *p*-value: <0.001

Model 3: gexp ~ time + treat

Coefficients:

	Estimate	Std. Error	*t* value	*p*-value
(Intercept)	4.93265	0.36292	13.592	<0.001
time	0.64490	0.03733	17.274	<0.001
treatA21	0.90969	0.45494	2.000	0.047
treatA24	0.53456	0.45494	1.175	0.241
treatB18	-0.19151	0.45494	-0.421	0.674
treatB21	1.12807	0.45494	2.480	0.014
treatB24	0.63681	0.45494	1.400	0.163
treatC18	0.29556	0.45494	0.650	0.516
treatC21	0.34120	0.45494	0.750	0.454
treatC24	1.55996	0.45494	3.429	<0.001

Residual standard error: 1.762 on 260 degrees of freedom
Multiple R-squared: 0.5533, Adjusted R-squared: 0.5378
MS ratio: 35.78 on 9 and 260 df, *p*-value: < 2.2e-16

Question continued on the next page

Gene expression responses:

Top row – variety A, middle row – variety B, bottom row – variety C

Left column – 18°C, centre column – 21°C, right column – 24°C