EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2010

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 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. An experiment is to be conducted involving five treatments A – E, and there are enough units available to replicate each treatment five times. However, the experimenter can only deal with five units each day and therefore intends to spend five days on the experiment. There may be systematic differences between days, and also differences due to the order in which treatments are carried out each day.

(i) Explain how a Latin square design can be used to eliminate systematic variation, and write down the linear model that is the basis for analysing data from this experiment, stating the properties of each term in it.

(3)

(ii) The experimenter tries to write down a plan for the week's work, and asks you how to construct the necessary design. You show him a table of standard 5×5 Latin squares and he says some of those look rather "systematic". Explain carefully how to choose a square at random from all possible 5×5 Latin squares.

(3)

(iii) The experiment is finally carried out using the following plan, which also shows the measurement *y* obtained from each unit.

Order	*Day 1*	*Day 2*	*Day 3*	*Day 4*	*Day 5*
First	E: 8.21	D: 5.64	A: 6.52	B: 8.80	C: 6.18
Second	B: 8.55	E: 9.12	C: 5.97	D: 6.63	A: 6.96
Third	D: 6.06	A: 7.37	E: 8.58	C: 6.11	B: 8.95
Fourth	C: 6.34	B: 8.82	D: 6.70	A: 6.85	E: 8.87
Fifth	A: 7.89	C: 6.84	B: 9.03	E: 9.31	D: 6.85

$\Sigma y^2 = 1436.4189$, grand total = 187.15.

Treatment totals are: A: 35.59; B: 44.15; C: 31.44; D: 31.88; E: 44.09.

(a) Construct the analysis of variance for these data.

(10)

(b) Find a 95% confidence interval for the difference between the means, $\bar{y}$, for treatments D and E.

(3)

(c) Comment briefly on the Day and Order terms in the analysis.

(1)

2. (i) A factorial experiment is to be carried out using five factors, A – E, each at two levels. The scientist in charge believes that up to 64 units (plots) can be found, but they will need to be grouped into blocks of 8 for operational reasons.

(a) Construct a suitable design for this experiment and write down the contents of the principal block. From this, write down also the contents of the other blocks.

(6)

(b) Write down the outline of the analysis of variance, listing the terms in it and their degrees of freedom.

(3)

(ii) After thinking further about the experiment, the scientist decides that a sixth factor F could usefully be included.

(a) Construct a suitable design for this version of the experiment. [There is no need to list the contents of the blocks this time.]

(4)

(b) Write down the outline of the analysis of variance, listing the terms in it and their degrees of freedom.

(4)

(iii) Comment on which of (i) and (ii) you think would lead to the better analysis, giving reasons for your answer.

(3)

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

Write down the conditions for this contrast to be *orthogonal* to a second contrast between the same totals $\{T_i\}$.

Explain why contrasts should be chosen to be mutually orthogonal if this is possible.

(5)

(b) In an agricultural experiment, several methods of controlling a pest which attacks a crop are being examined. An untreated "control" O is included to check whether there is any infestation present on the site of the experiment. Treatment S is a standard, which has been used on the site previously, while A, B, C, D are experimental treatments that are new (and possibly improved) methods of controlling the pest. A and B use one compound in different physical forms, while C and D use another compound applied at different strengths. Each of these six treatments is replicated five times in a completely randomised layout.

(i) Construct a set of five orthogonal contrasts which answer useful questions about the performance of the treatments. Explain what each contrast measures.

(5)

(ii) The totals $\{T_i\}$ for the 5 replicates of O, S, A, B, C, D respectively are 45, 55, 70, 72, 63, 87; the total corrected sum of squares is 432.8 and the corrected sum of squares for treatments (with 5 d.f.) is 212.267.

Construct an analysis of variance, splitting the sum of squares for treatments into the contrasts specified in (i).

(8)

Comment on the results of this analysis.

(2)

4. A scientist believes that the concentration of a chemical in the root cells of a particular species of plant is linearly related to the quantity of an added soil nutrient. An experiment has been carried out using a completely randomised design, with four equally spaced levels of the nutrient and three varieties of the plant species. Two replicates were made, giving 24 observations altogether. The totals of the two replicates were as follows.

Nutrient Level	*1*	*2*	*3*	*4*
Variety X	34	39	44	38
Y	26	37	41	35
Z	27	24	41	48

The sum of the squares of the 24 observations was 8202, and the coefficients of the linear component of the nutrient level factor are (–3, –1, 1, 3).

(i) Construct an analysis of variance and carry out any appropriate tests in order to check the scientist's belief.

(15)

(ii) Draw a graph of the 12 variety/level totals (or means) and comment on any inferences that might be drawn from it.

(4)

(iii) Suggest any follow-up study that the scientist should make.

(1)

5. (i) 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. Give an example of a relationship between a single response variable y and a single explanatory variable x that cannot be expressed as a linear model. (4)

(ii) A physicist has carried out an experiment to investigate the effect of temperature on electrical resistivity using three different alloys P, Q and R. Measurements were taken at nine different temperatures for each of the alloys.

Analyses have been carried out by fitting various models using x_1 as temperature and x_2 and x_3 representing the three alloys, with coding as follows.

Alloy	x_2	x_3
P	0	0
Q	1	0
R	0	1

The regression sums of squares for some of the fitted models are as follows.

Model	*Variables in model*	*Regression sum of squares*
A	x_1	2847
B	x_1, x_2, x_3	35232
C	x_1, x_1x_2, x_1x_3	33800
D	$x_1, x_2, x_3, x_1x_2, x_1x_3$	35517

The corrected total sum of squares is 35 673.

(a) Explain why all of these models are *linear models*. (1)

(b) The physicist remembers that one of these models represents parallel straight lines with a different intercept for each alloy. Which model is this? Justify your answer. (1)

(c) Describe what is being modelled in the other three models. (3)

(d) Based on the information given in the table, which model would you suggest for describing the data? Justify your answer. (5)

(e) The physicist says that the R-squared value for model D is in excess of 99% and that therefore this model must be correct. Verify the value of R-squared and comment on the physicist's statement. (2)

(f) Someone else states that, based on research they have read about, they believe the model is inadequate as it does not allow for curvature. What advice would you give about how to investigate other models? (4)

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

(4)

By considering the linear regression of a dependent variable *y* on a single predictor variable *x*, sketch scatter plots demonstrating data points with

(i) high leverage and high influence,

(ii) high leverage and low influence,

(iii) low leverage and low influence.

(3)

In order to investigate factors affecting the cost of a service, a regression analysis with three predictor variables has been carried out. Some of the regression diagnostics produced by a computer package are given in the table below. The cases are numbered in the order in which they were observed.

Unfortunately there is only one copy of the output and this has been stained. The asterisks denote values that are illegible on the printout.

Case number	*Observed cost*	*Fitted cost*	*Residual*	*Standardised residual*	*Leverage*	*Cook's distance*
1	13320	8817	4503	2.39	*	0.88
2	2850	1246	1604	1.06	*	0.38
3	5580	2900	2680	1.62	0.60	0.66
4	5055	5944	–889	–0.47	*	0.03
5	6060	7956	–1896	–1.01	0.48	0.16
6	4800	5553	–753	–0.34	0.27	0.01
7	4290	4410	–120	–0.07	0.61	0.00
8	4875	5720	–845	–0.42	0.41	0.02
9	3018	4686	–1668	–0.73	0.23	0.03
10	3180	4473	–1293	–0.87	0.68	0.26
11	4140	4933	–793	–0.43	0.49	0.03
12	3339	3868	–529	–0.32	0.60	0.03

Use the information in the table to estimate the residual standard deviation and hence the three missing leverage values.

(4)

Interpret the information in the table and hence identify possible problems with the fitted model.

(9)

7. In a social study in a large city in England, 300 males and 300 females were randomly selected from two different areas. They were each asked whether or not they were satisfied with their present jobs. The results were as follows.

	Male		**Female**	
	Area 1	*Area 2*	*Area 1*	*Area 2*
Satisfied	126	53	86	64
Not satisfied	95	26	92	58

A log-linear model has been proposed for these data, the full model being

$$\log(\lambda_{ijk}) = \mu + S_i + M_j + A_k + (SM)_{ij} + (SA)_{ik} + (MA)_{jk} + (SMA)_{ijk},$$

where the observation in cell *ijk* has a Poisson distribution with mean λ_{ijk}, independently of observations in the other cells, *S* is the satisfaction factor, *M* is the sex factor and *A* is the area factor.

(i) Interpret, in terms understandable to a non-statistician, the terms $(SM)_{ij}$ and $(SMA)_{ijk}$.

(2)

(ii) Part of the computer output on fitting the model with only the first order effects is shown below.

```
------------------------------------
    number |      Coef.    Std. Err.
-----------+------------------------
 satisfied |    0.1939       0.0820
     area2 |   -0.6857       0.0865
    female |    1.62e-14     0.0816
      cons |    4.5010       0.0787
------------------------------------
```

(a) Explain why the coefficient for sex must be zero for this dataset, but why it is necessary in the model.

(2)

Question 7 is continued on the next page

(b) Another person has also analysed the data using a log-linear model whose systematic component includes only the first order effects, but has obtained the following output.

Factor	*Level*	*Coefficient*
Constant		4.2551
S	Satisfied	0.0970
	Not satisfied	–0.0970
M	Male	0.0000
	Female	0.0000
A	Area 1	0.3428
	Area 2	–0.3428

By considering satisfied males in area 1 show that the two sets of output give the same linear predictor score, and explain what the two analysts have done differently to get the different first order estimates. Hence show that the models are equivalent.

(6)

(iii) Various log-linear models have been fitted, using consistent parametrisation, but including different terms. The results are summarised in the table.

Terms in model	*Scaled Deviance*
μ	94.77
μ, *S*, *M*, *A*	22.57
μ, *S*, *M*, *A*, *SM*	16.91
μ, *S*, *M*, *A*, *SA*	21.18
μ, *S*, *M*, *A*, *AM*	8.66
μ, *S*, *M*, *A*, *SM*, *SA*	15.51
μ, *S*, *M*, *A*, *SM*, *AM*	2.99
μ, *S*, *M*, *A*, *AM*, *SA*	7.27
μ, *S*, *M*, *A*, *SM*, *SA*, *AM*	0.53

(a) What would be the value of the scaled deviance for the saturated model which includes all three main factors, all three two-way interactions and the three-way interaction?

(1)

(b) Using backward elimination, obtain the best model for the data in terms of fit and parsimony, showing all your reasoning.

(6)

(c) Which factors appear to affect whether respondents are satisfied with their present jobs?

(3)

8. A company is interested in investigating whether either the age or sex of a person, or both, has any effect on whether or not the person is prepared to take part in a survey planned by the company. Their market researcher has carried out a pilot study, the results of which are given in the table below. The data are in the form r/n where r is the number of people saying that they are prepared to take part out of a total of n people asked.

	Age < 30	*30 ≤ Age ≤ 59*	*Age > 59*
Male	30/52	18/42	8/19
Female	27/56	28/51	19/37

A logistic regression model has been proposed for these data using the variable x_1 for the sex of the individual (coded 0 for male and 1 for female) and x_2, x_3 as dummy variables representing the three age groups, coded as shown below.

Age group	x_2	x_3
Age < 30	0	0
30 ≤ Age ≤ 59	1	0
Age > 59	0	1

Part of the output from a logistic regression program is given below.

```
---------------------------------
       |    Coef.      Std. Err.
-------+-------------------------
   X1  |   0.0884        0.2532
   X2  |  -0.1354        0.2833
   X3  |  -0.1953        0.3317
  cons |   0.0655        0.2331
---------------------------------
```

(i) Write down the form of the logistic regression model for these data. Make sure that you define the logit transformation.

(3)

(ii) Someone has suggested that it would be better to code age using a single variable with values 0, 1 and 2. What would your response be to this suggestion?

(3)

Question 8 is continued on the next page

(iii) One manager wants to know whether any of the coefficients are statistically significant, whereas another wants 95% confidence intervals for the model parameters.

(a) Discuss whether any of the coefficients are statistically significant.

(2)

(b) Compute an approximate 95% confidence interval for the coefficient of sex in the linear predictor.

(2)

(c) Given that this is a pilot study, would you recommend that the managers concentrate on the statistical significance or the 95% confidence intervals at this stage? Justify your answer.

(2)

(iv) Estimate from the model the proportion of females aged more than 59 who would be prepared to take part in the survey.

(4)

(v) One of the managers has heard about odds ratios and wants to know how the values in the output shown relate to odds ratios. Explain how the coefficient 0.0884 is related to an odds ratio and how it is interpreted.

(4)

Note

The version of this paper that was worked in the actual examination contained two minor errors.

In question 3(b) part (ii), the first data item (45) was accidentally printed as 40. In question 4, in the data table the entry for Y level 2 (37) was accidentally printed as 27. The Society apologises for these errors.

Full credit was of course given to candidates for working the questions as printed.