ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2016

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 16 printed pages.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

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

(3)

(ii) Describe the concept of *multicollinearity*. Explain why a correlation matrix showing pairwise correlations between all pairs of potential explanatory variables may not provide clear information about multicollinearity.

(4)

In a forestry study, data were collected from a random sample of 44 loblolly pine trees with the objective of describing average crown width (CW – the maximum width of the foliage growing out of the trunk) in terms of explanatory variables that are simple to measure including diameter at breast height (DBH), total tree height (Ht), height to live crown (HLC), average branch diameter (DiamB) and average branching angle (AngB).

Multiple linear regression has been applied to determine whether there is a useful relationship. The analysis output **on the next page** shows the pairwise correlation coefficients for the six variables, the summary of a backwards elimination variable selection process where terms are eliminated from the model if the variance ratio is less than 1.00, and the fitted parameters for the model finally selected using the backwards elimination process.

(iii) Briefly describe how stepwise variable selection approaches work, considering both the *forwards selection* and *backwards 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)

(iv) Interpret the results from the backwards elimination stepwise process, including an explanation of why the total height (Ht) predictor variable has been omitted from the final model, even though it has the second strongest correlation with average crown width (CW).

(5)

(v) Identify the assumptions associated with the multiple linear regression model, and describe how you would assess these assumptions informally.

(4)

Output for Question 1 is on the next page

Correlation matrix

DBH	-0.0672				
DiamB	0.4263	0.2334			
HLC	0.1727	0.0167	0.0483		
Ht	0.3170	0.5056	0.4544	0.1999	
CW	-0.1120	0.5762	0.1795	-0.1280	0.3683
	AngB	DBH	DiamB	HLC	Ht

Backwards Elimination 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:   0.07430 - Dropping DiamB          0.07563 - Dropping AngB
          0.07597 - No change               0.07627 - Dropping Ht
          0.07663 - Dropping HLC            0.09154 - Dropping DBH
Chosen action: Dropping DiamB

Step 2:   0.07379 - Dropping AngB           0.07430 - No change
          0.07514 - Dropping HLC            0.07525 - Dropping Ht
          0.09012 - Dropping DBH
Chosen action: Dropping AngB

Step 3:   0.07374 - Dropping Ht             0.07379 - No change
          0.07499 - Dropping HLC            0.09340 - Dropping DBH
Chosen action: Dropping Ht

Step 4:   0.07374 - No change               0.07413 - Dropping HLC
          0.10916 - Dropping DBH
Chosen action: no change
```

Accumulated analysis of variance

	df	SS	MS	Variance ratio	*p*-value
Full model	5	1.80913	0.36183	4.76	0.002
Residual	39	2.96287	0.07597		
Total	44	4.77200	0.10845		
Step 1 – starting with Full model					
Dropping DiamB	-1	-0.00924	0.00924	0.12	0.729
Step 2 – starting with model after Dropping DiamB					
Dropping AngB	-1	-0.05310	0.05310	0.70	0.408
Step 3 – starting with model after Dropping DiamB and AngB					
Dropping Ht	-1	-0.07196	0.07196	0.95	0.336

Final model: Constant + DBH + HLC

Parameters for chosen model

Parameter	estimate	s.e.	t value
Constant	0.412	0.522	0.79
DBH	0.2150	0.0462	4.65
HLC	-0.0320	0.0289	-1.11

2. An agricultural experiment was conducted to investigate the effect of sulphur fertilizer on the grain yield (tonnes/ha) of spring barley. Five levels of sulphur fertilizer were applied (0, 10, 20, 30 and 40 kg). The field in which the experiment was sited has a gradual slope so that plots at the top of the slope tend to have lower moisture levels than those at the bottom, and, in a previous experiment, strips of land from the top to the bottom of the field had received different levels of nitrogen fertilizer. This experiment was therefore arranged following a Latin square design.

(i) Explain how a Latin square design allows the elimination of any systematic variation between different positions from the top to the bottom of the field ("rows"), or across the field relating to the previous nitrogen fertilizer treatments ("columns"). (3)

The table below gives the total grain yields (tonnes/ha) for the five plots receiving each level of sulphur fertilizer, plus the overall total.

Sulphur fertilizer (kg)					Total
0	10	20	30	40	
22.92	24.12	24.98	24.20	22.76	118.98

The five column totals (for five plots each) are 24.58, 23.70, 24.37, 23.47 and 22.86, and the five row totals (again for five plots each) are 21.39, 22.02, 24.35, 24.68 and 26.54, and the sum of squares for the 25 observations is 571.69.

(ii) Construct an analysis of variance to assess the effect of sulphur fertilizer, taking full account of the design of the experiment, and comment on the results, including the importance of allowing for systematic variation in the field. (7)

(iii) Construct a 95% confidence interval for the difference in mean grain yield between sulphur fertilizer levels of 0 and 10 kg. (2)

(iv) Explain why the use of orthogonal contrasts is particularly helpful in exploring sources of variability within analysis of variance, and state how the coefficients for any two contrasts can be used to check whether the contrasts are orthogonal. (3)

The orthogonal linear and quadratic polynomial contrasts for five equally spaced factor levels have coefficients $(-2, -1, 0, +1, +2)$ and $(+2, -1, -2, -1, +2)$ respectively.

(v) Calculate the contrast values and the associated sums of squares for these orthogonal linear and quadratic polynomial contrasts for the effect of sulphur fertilizer in this experiment. Interpret the results in terms of the shape of the response to level of sulphur fertilizer. Indicate whether there is any evidence for lack of fit for whichever of the linear or quadratic models you determine to be the most appropriate for describing the mean grain yields. (5)

3. An experiment is to be conducted to compare the responses of nine treatments A – I. There is sufficient material for eight replicate samples to be processed using each of the nine treatments and up to twelve 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 eight samples, batches of samples from this initial step can be stored for processing on subsequent days.

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

(4)

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

$$rt = bk \qquad \lambda(t-1) = r(k-1)$$

Here, t is the number of treatments, r is the number of replicates of each treatment, 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) Using the relationships in part (ii), identify the block sizes, k, that could result in balanced incomplete block designs making complete use of the available eight replicates per treatment, for the cases

(a) $\lambda = 1$,

(b) $\lambda = 2$,

(c) $\lambda = 5$.

For each of these cases also give the number of blocks, b. For the case of $\lambda = 4$, show that a balanced incomplete block design does not exist for nine treatments.

(5)

(iv) For eight replicates of nine treatments, the balanced incomplete block design with $\lambda = 3$ requires 18 blocks each containing four units. Construct the allocation of treatments to blocks for this design, describing the process that you have followed to produce this allocation. [Hint: the design can be split into two sets of nine blocks, each set containing four replicates of each treatment, and constructed using a cyclic approach.]

(6)

(v) Describe how this balanced incomplete block design should be randomised to be performed on six days with three blocks being processed per day.

(3)

4. (i) Discuss the differences between linear and non-linear regression models, and briefly describe how the Newton-Raphson procedure is used to find the optimal parameter values for a non-linear model.

(6)

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

(a) $y = a + bx + cx^2 + dx^3$ [parameters a, b, c, d]

(b) $y = a + c\exp(-\exp(-b(x-m)))$ [parameters a, b, c, m]

(c) $y = a + be^{-cx} + dx$ [parameters a, b, c, d]

(d) $y = a + b\sin\left(\frac{2\pi x}{w}\right) + c\cos\left(\frac{2\pi x}{w}\right)$ [parameters a, b, c, w]

(4)

An experiment was conducted to investigate the behaviour of genes associated with leaf development in plants. Forty-four plants were grown in a glasshouse, with four plants randomly allocated to each of 11 sampling occasions (every two days from day 19 to day 39 after sowing), and destructively tested. A response variable referred to as "Expression" was measured for various genes for each of the 44 plants. The graph **on the next page** shows the Expression data for one particular gene plotted against days after sowing. The computer output **on the next page** shows the results of fitting a logistic (sigmoidal growth) curve to these data, the model being

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

where y is the response variable (Expression), x represents the explanatory variable (the number of days after sowing) and a, b, c and m are parameters.

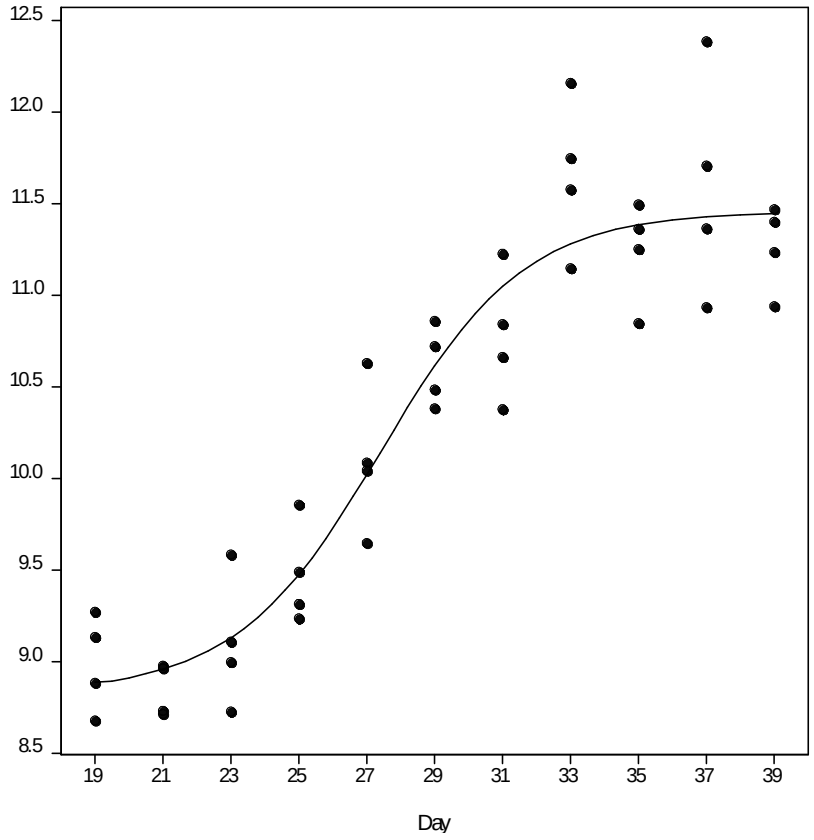
(iii) Interpret the results of the analysis, including a description of how the fitted parameters relate to the Expression response over time. Estimate the time at which an Expression response of 10.5 is reached.

(7)

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

(3)

Output for Question 4 is on the next page



Nonlinear regression analysis

```
Response variate:  Expression
Explanatory:       Day
Fitted Curve:      a + c/(1 + exp(-b*(x - m)))
```

Summary of analysis

Source	df	SS	MS	Variance ratio
Regression	3	45.167	15.0556	114.23
Residual	40	5.272	0.1318	
Total	43	50.439	1.1730	

```
Percentage variance accounted for 88.8
Standard error of observations is estimated to be 0.363
```

Estimates of parameters

Parameter	estimate	s.e.
b	0.472	0.116
m	27.415	0.551
c	2.617	0.242
a	8.839	0.175

BLANK PAGE

5. In an industrial study, failure rates of a mechanical component were tested in three independent laboratories under pre-specified conditions. Batches of the mechanical components were obtained from three different manufacturers (labelled A, B and C), and the components were tested at four different temperatures (0, 5, 10 and 15 degrees C) for four durations (1, 2, 5 and 10 hours). A sample of 20 components from each manufacturer was tested under each combination of temperature and duration by each of the three independent laboratories.

To analyse these data to assess for differences in failure rates between manufacturers, at different temperatures and over different testing durations, and for combinations of these factors, the study director first converted each failure count into a percentage failure rate. The output from an analysis of variance of the percentage failure rate data is shown **on the next two pages**, together with the associated residual plots.

(i) Interpret the analysis of variance table, providing a clear description of how percentage failure changes in response to each of the terms included in the fitted model, using language that would be understandable to a non-statistician.

(7)

(ii) Identify the assumptions associated with this analysis. Describe how the residual plots have been constructed and how they can be used to assess the validity of these assumptions.

(5)

(iii) Describe how these data, as percentages of the fixed number of components being tested for each treatment combination, might cause one or more of these assumptions to be invalid. Identify how such a violation of the assumptions might be anticipated to be revealed in residual plots. Discuss briefly the extent to which the residual plots presented here indicate such a violation.

(5)

(iv) Identify how this issue might be overcome within the framework of analysis of variance. Describe briefly an alternative analysis approach that could be applied to take full account of the form of the data.

(3)

Output for Question 5 is on the next two pages

Analysis of variance

Source of variation	df	SS	MS	Variance ratio	*p*-value
laboratory	2	415.6	207.8	1.97	0.145
duration	3	17851.9	5950.6	56.50	<.001
manufacturer	2	12213.5	6106.8	57.98	<.001
temperature	3	9700.5	3233.5	30.70	<.001
duration.manufacturer	6	2772.6	462.1	4.39	<.001
duration.temperature	9	814.1	90.5	0.86	0.565
manufacturer.temperature	6	1644.8	274.1	2.60	0.022
duration.manufacturer.temperature	18	1646.9	91.5	0.87	0.616
Residual	94	9901.0	105.3		
Total	143	56960.9			

Tables of means

duration	1	2	5	10
	35.42	43.06	47.92	65.69

manufacturer	A	B	C
	41.77	61.04	41.25

temperature	0	5	10	15
	36.11	45.00	53.61	57.36

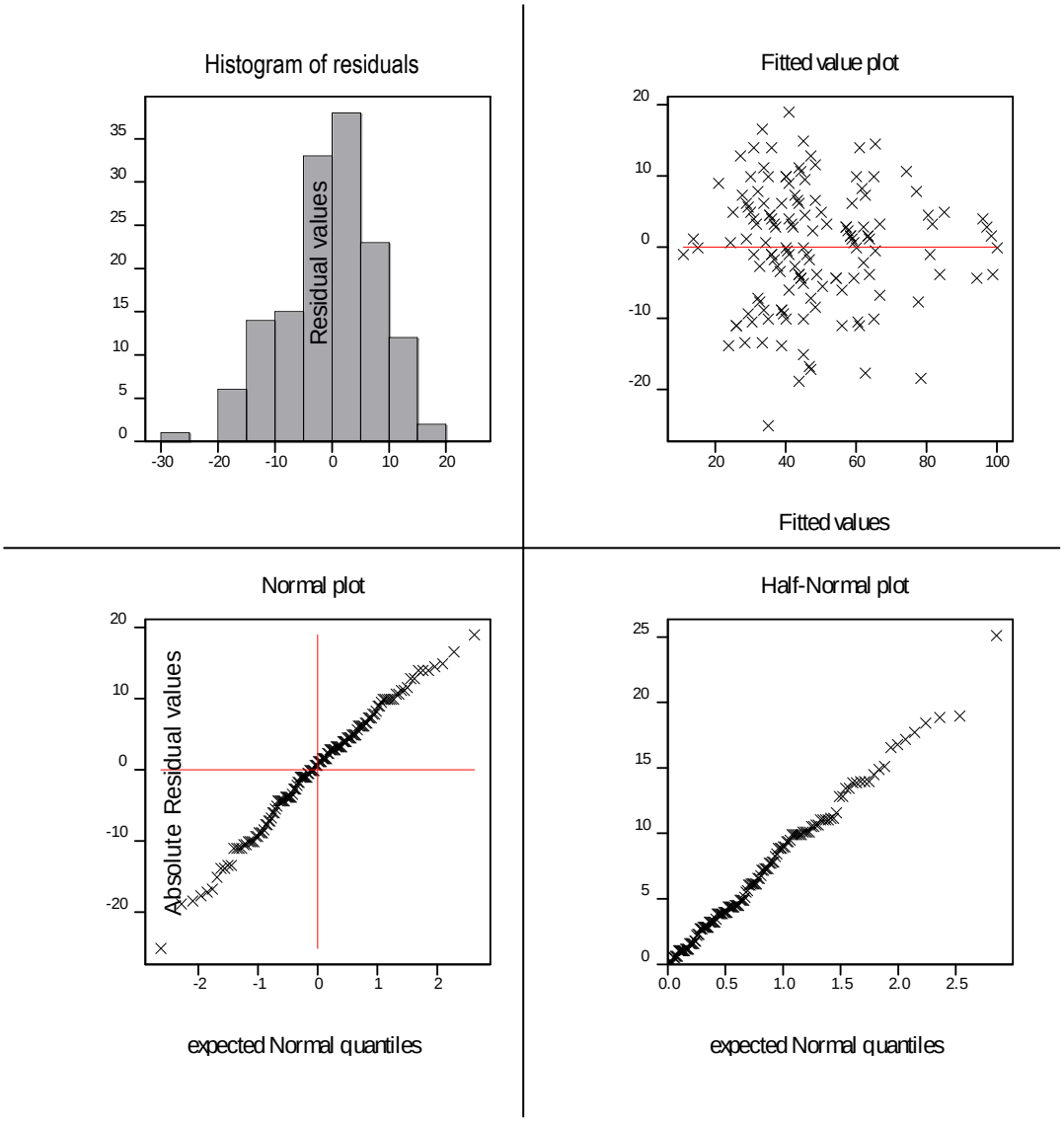
	manufacturer		
duration	A	B	C
1	32.50	44.17	29.58
2	37.50	49.17	42.50
5	38.33	65.42	40.00
10	58.75	85.42	52.92

	temperature			
duration	0	5	10	15
1	24.44	33.89	36.67	46.67
2	30.56	35.56	54.44	51.67
5	36.11	47.22	51.67	56.67
10	53.33	63.33	71.67	74.44

	temperature			
manufacturer	0	5	10	15
A	33.33	39.58	44.17	50.00
B	42.08	56.67	70.42	75.00
C	32.92	38.75	46.25	47.08

Standard errors of differences of means

Table	duration	manufacturer	temperature	duration manufacturer	duration temperature	manufacturer temperature
rep.	36	48	36	12	9	12
s.e.d.	2.419	2.095	2.419	4.190	4.838	4.190


Histogram of residuals
Residual values
35
30
25
20
15
10
5
0
-30
-20
-10
0
10
20
Fitted value plot
20
10
0
-10
-20
20
40
60
80
100
Fitted values
Normal plot
Absolute Residual values
20
10
0
-10
-20
-2
-1
0
1
2
expected Normal quantiles
Half-Normal plot
25
20
15
10
5
0
0.0
0.5
1.0
1.5
2.0
2.5
expected Normal quantiles

6. A laboratory experiment is to be performed to assess the impact of four chemical components, A – D, all at two levels, and two temperature factors, E and F, each also at two levels, on the strength of a plastic. The available resources allow for 32 treatment combinations to be assessed, but with a maximum of 8 runs of the process possible per day, and with anticipated differences between the responses observed on different days. It is believed that there will be interactions between all pairs of chemicals, and also between each chemical component and each of the temperature factors, but not between the two temperature factors.

(i) Briefly describe the concepts of a *confounded factorial design* and a *fractional factorial design*, including why these two concepts might be combined.

(5)

(ii) Identify appropriate confounding and defining (fractionating) contrasts to generate a confounded half-replicate design for this experiment in four blocks of size 8, so that all main effects and appropriate two-factor interactions can be estimated. Explain your choices of confounding and defining contrasts.

(5)

(iii) Construct the allocation of treatment combinations to blocks for the principal half fraction.

(6)

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

(4)

7. An experiment was conducted to assess the efficacy of two species of insect predators (melanarius, tachyporus) in controlling aphids (an insect pest) on plants grown in separate insect-proof cages. Observations were made over eight consecutive week-long periods, with fresh plants, aphid populations and predators being introduced each week. Eight plants were observed each week, with two plants receiving each of four treatment combinations (no predators, melanarius only, tachyporus only, both predators). In four of the eight weeks a fungicide was also applied to all the plants, to control outbreaks of mildew. It was expected that there might be differences in the development of aphid populations and the efficacy of the predators between the week-long periods because of variation in the temperature as well as the use of different batches of predators and aphids.

At the end of each week period, a count of the number of aphids was made for each plant. The output **on the next page** shows the results of analysing the counts of aphids at the end of each week using a generalised linear model which allows overdispersion to be detected if it is present. The application of the fungicide to all the plants in particular weeks potentially imposes a more complex design structure, but a simplified structure was assumed to allow analysis using a generalised linear model. The fitted model allows for the main effects of each predator and the fungicide treatment, and for the interactions between these effects. The fitted model also includes terms accounting for differences between periods after allowing for fungicide effects, and differences in the separate and combined effects of the two predators between periods (again, after allowing for fungicide effects), these potentially contributing extra variability beyond the plant-to-plant variability against which treatment effects should be assessed.

(i) Identify the components of a generalised linear model (GLM), showing how the expected value of the response is related to the explanatory model. Describe the particular form of GLM that is appropriate for these data.

(5)

(ii) Briefly describe the method of iterative reweighted least squares used in fitting a generalised linear model.

(4)

(iii) Interpret the results of the analysis, in particular identifying the separate and combined efficacies of the two insect predators, and the impact of the fungicide application.

(8)

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

(3)

Output for Question 7 is on the next page

Accumulated analysis of deviance

Change	d.f.	deviance	mean deviance	deviance ratio	*p*-value
+ fungicide	1	23.91	23.91	1.88	0.179
+ tachyporus	1	990.58	990.58	78.08	<.001
+ melanarius	1	0.88	0.88	0.07	0.794
+ fungicide.period	6	1116.41	186.07	14.67	<.001
+ fungicide.tachyporus	1	183.47	183.47	14.46	<.001
+ fungicide.melanarius	1	4.63	4.63	0.36	0.550
+ tachyporus.melanarius	1	3.13	3.13	0.25	0.623
+ fungicide.tachyporus.period	6	516.48	86.08	6.79	<.001
+ fungicide.melanarius.period	6	108.37	18.06	1.42	0.236
+ fungicide.tachyporus.melanarius	1	0.01	0.01	0.00	0.975
+ fungicide.tachyporus.melanarius.period	6	197.20	32.87	2.59	0.037
Residual	32	405.97	12.69		
Total	63	3551.05	56.37		

Dispersion parameter is estimated to be 12.7 from the residual deviance

The parameters estimate the effect of applying the fungicide or each predator

Estimates of some parameters

Parameter	estimate	s.e.	t value	*p*-value	antilog of estimate
Constant	5.218	0.185	28.14	<.001	184.5
fungicide yes	–0.455	0.298	–1.53	0.136	0.6341
tachyporus yes	–1.140	0.377	–3.03	0.005	0.3198
melanarius yes	0.108	0.255	0.42	0.676	1.114
fungicide yes.tachyporus yes	0.092	0.592	0.15	0.878	1.096
fungicide yes.melanarius yes	–0.099	0.416	–0.24	0.813	0.9055
tachyporus yes.melanarius yes	0.425	0.486	0.87	0.388	1.529
fungicide yes.tachyporus yes.melanarius yes	0.458	0.750	0.61	0.545	1.581

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

Predicted means (standard errors)

		melanarius	
		No	Yes
tachyporus	No	128.34 (14.15)	131.19 (14.57)
	Yes	43.81 (8.18)	69.63 (11.27)

		tachyporus	
		No	Yes
fungicide	No	97.00 (11.11)	13.03 (2.85)
	Yes	80.66 (9.72)	42.66 (6.96)

		melanarius	
		No	Yes
fungicide	No	54.78 (8.16)	55.25 (8.06)
	Yes	72.62 (9.71)	50.69 (6.97)

8. (i) Distinguish between *fixed effects* and *random effects* in modelling experimental data. Use examples to illustrate when model terms should be identified as fixed effects, and when as random effects.

(4)

In an environmental study, a water company is assessing the efficacy of four different types of water treatment plants on the down-stream water quality. For each of the types of water treatment plant, inspectors have selected three locations at random from within the river catchment the company is responsible for. Water samples have then been collected from each location on each of five dates over a three-month period on a day after it has rained in the local area, and an overall measure of water quality has been obtained for each sample. Variation in the water quality values could therefore be due to differences between the different types of water treatment plant, differences between the locations for each type of water treatment plant, and differences between the sampling occasions at each location.

(ii) The model suggested for analysing these data is

$$y_{ijk} = \mu + t_i + l_{ij} + \varepsilon_{ijk} .$$

Identify what each of the terms (y, μ, t, l and ε) in the model represents, and the ranges for the three indices (i, j and k), and state clearly the assumptions needed to conduct an analysis of this model.

(6)

(iii) A partially completed analysis of variance for these data is as follows.

Source of variation	Sum of squares	Expected mean square
Between types	2562.82	$\sigma^2 + 5\sigma_l^2 + 5\Sigma t_i^2$
Between locations within types	404.00	$\sigma^2 + 5\sigma_l^2$
Between samples within locations	208.31	σ^2

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

(10)

BLANK PAGE