ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2015

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) Describe what is meant by a *balanced incomplete block design*, including in your description how the following relationships are used to identify the structure of a balanced incomplete block design given the number of treatments, t, and the number of times, λ, that each pair of treatments appear together in a block.

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

Here, r is the number of replicates of each treatment, b is the number of blocks, and k is the number of units per block.

(6)

In a laboratory experiment comparing the effects of eight treatments A – H, there is sufficient experimental material to allow for seven replicates of each treatment. However, the conditions under which the experiment is performed will not remain constant for the time taken for all eight treatments to be assessed.

(ii) Write down the maximum possible value of k, the block size, for a balanced incomplete block design for eight treatments ($t = 8$). Using the relationships in part (i), identify the possible block sizes, k, that result in balanced incomplete block designs for the cases

(a) $\lambda = 1$,

(b) $\lambda = 3$,

(c) $\lambda = 6$.

In each of these cases, give the numbers of blocks, b, and replicates, r, required for the design. For the case of $\lambda = 2$, show that a balanced incomplete block design does not exist for eight treatments.

(5)

(iii) Briefly discuss the relative advantages and disadvantages of the three balanced incomplete block designs identified in part (ii).

(3)

(iv) Suppose now that the experimental conditions will remain sufficiently constant for no more than five experimental units to be included in a block. Given this constraint, construct the allocation of treatments to blocks for the balanced incomplete block design identified in part (ii) with the largest possible block size.

(6)

2. In a field trial to investigate the efficacy of three insecticide treatments, one of the treatments is applied directly to the seeds prior to sowing (seed treatment), while the other two insecticides will be applied as sprays to plants in the field.

It is proposed to include six treatment combinations, as follows.

A	No seed treatment, no spray
B	No seed treatment, spray 1
C	No seed treatment, spray 2
D	Seed treatment, no spray
E	Seed treatment, spray 1
F	Seed treatment, spray 2

Up to 36 field plots are available, allowing for up to six replicates of each treatment combination. A natural infestation of pests is expected in the field containing the plots, but it is not known from which direction the pests might arrive. There is space to arrange the 36 field plots in an array of 6 rows by 6 columns.

(i) Identify the advantages and disadvantages of arranging this trial as

(a) a completely randomised design,

(b) a randomised complete block design,

(c) a Latin square design.

For each of these designs include a sketch to indicate how the plots would be arranged in the field, and show the dummy analysis of variance table.

(13)

(ii) Describe the randomisation process that should be followed for the allocation of treatments to field plots following a Latin square design.

(3)

(iii) The application of the insecticide sprays is most effective when applied to larger areas, so the experimenter proposes to group the 36 field plots into 18 pairs of adjacent plots. Each pair of plots would then receive a particular spray treatment (no spray, spray 1, spray 2), with one plot in each pair having the seed treatment applied and the other having no seed treatment. Describe a suitable field layout for this approach (including a sketch), and describe the randomisation process that should be followed for the allocation of treatments following this design.

(4)

3. (i) Briefly describe the concepts of a *confounded factorial design* and a *fractional factorial design*, including in your answer the different ways in which *defining (fractionating) contrasts* and *confounding contrasts* are used to construct such designs.

(6)

In an industrial experiment, the impacts of 7 two-level factors, A – G, on the production of an electronics component are to be assessed. The available resources allow for 32 runs of the production process, with a maximum of 8 runs possible within a single day, but differences in environmental conditions between days might influence the response. At this stage in the study, emphasis is on determining which of the factors have a major impact on the response, and whether there are any clear interactions between pairs of factors. The proposed approach is to use a quarter fraction of the full factorial design, with 8 runs per day on each of 4 days.

(ii) Identify two fractionating contrasts that would generate a quarter fraction of the full factorial design that allows estimation of all 7 main effects without aliasing these effects with any of the two-factor interactions. Explain how you have identified these fractionating contrasts.

(3)

(iii) For your chosen fractionating contrasts, identify the three terms that will be aliased with each of the main effects.

(3)

(iv) Construct the principal quarter fraction that your choice of fractionating contrasts produces, explaining how the treatment combinations included are identified.

(4)

(v) Identify two confounding contrasts that could be used to divide this quarter fraction into four blocks each containing 8 treatment combinations, ensuring that all 7 main effects are still estimable. Identify the 8 treatment combinations from part (iv) that will be allocated to each block.

(4)

4. Data were collected in 41 US cities on levels of air pollution (Y = annual mean concentration of sulphur dioxide in micrograms per cubic metre) and six potential explanatory variables for such pollution (*X1* = average annual temperature in degrees Fahrenheit; *X2* = number of manufacturing enterprises employing 20 or more workers; *X3* = population size in thousands; *X4* = average annual wind speed in miles per hour; *X5* = average annual rainfall in inches; *X6* = average number of days with rainfall per year).

Multiple linear regression was applied to these data, using an all-subset selection approach to identify the important variables associated with variation in levels of air pollution. Results from the different subset selections are shown **on the next two pages**, together with information on the parameter estimates when all six explanatory variables were included, and a scatter plot matrix showing the relationships among the seven variables.

(i) Explain the concepts of *leverage* and *influence*. For the case of a simple one-variable linear regression model, describe how the leverage values are obtained from the values of the explanatory variable, and give sketch plots to illustrate observations with

(a) high leverage but low influence,

(b) low leverage but high influence,

(c) high leverage and high influence,

giving brief descriptions of how your sketch plots illustrate these features. For the case of the air pollution data, more detailed analysis output identified three observations as having high leverage. Use the scatter plot matrix to identify the explanatory variables most likely to be associated with these three high leverage observations.

(6)

(ii) Describe the difference between the *coefficient of determination (R^2 statistic)* and the *adjusted coefficient of determination (adjusted R^2 statistic)*. Explain why the former will always increase as more variables are added to a multiple linear regression, while the latter should reach a maximum value that suggests a parsimonious model that fits the data well.

(4)

(iii) Explain how the Mallows' C_p statistic can be used to provide a graphical approach for choosing the 'best' model in an all-subset selection approach. Interpret the presented information on this statistic to identify the most appropriate model.

(5)

(iv) For the air pollution data, interpret the fitted model that contains all six explanatory variables in non-technical language understandable by the scientists who collected the data. Include an explanation for any variables that appear to have parameter estimates with the wrong signs, given the apparent relationships shown in the scatter plot matrix.

(5)

Output for Question 4 is on the next two pages

All possible subset selection

** indicates that the term is included in the model.*
– indicates that the term is not included.

Best subsets with 1 term

Adjusted R^2	C_p	Df	X1	X2	X3	X4	X5	X6
40.07	23.11	2	–	*	–	–	–	–
22.44	40.79	2	–	–	*	–	–	–
16.72	46.54	2	*	–	–	–	–	–
11.44	51.83	2	–	–	–	–	–	*
<0.00	64.96	2	–	–	–	*	–	–
<0.00	65.57	2	–	–	–	–	*	–

Best subsets with 2 terms

Adjusted R^2	C_p	Df	X1	X2	X3	X4	X5	X6
56.45	7.56	3	–	*	*	–	–	–
49.06	14.78	3	*	*	–	–	–	–
47.17	16.63	3	–	*	–	–	–	*
39.09	24.53	3	–	*	–	–	*	–
38.88	24.73	3	–	*	–	*	–	–
37.54	26.05	3	*	–	*	–	–	–
33.23	30.26	3	–	–	*	–	–	*
20.88	42.33	3	–	–	*	–	*	–

Best subsets with 3 terms

Adjusted R^2	C_p	Df	X1	X2	X3	X4	X5	X6
58.64	6.36	4	–	*	*	–	–	*
58.11	6.86	4	*	*	*	–	–	–
56.01	8.87	4	–	*	*	–	*	–
56.00	8.87	4	–	*	*	*	–	–
52.67	12.04	4	*	*	–	–	*	–
50.84	13.79	4	*	*	–	–	–	*
50.83	13.79	4	*	*	–	*	–	–
46.85	17.58	4	–	*	–	*	–	*

Best subsets with 4 terms

Adjusted R^2	C_p	Df	X1	X2	X3	X4	X5	X6
59.96	6.07	5	*	*	*	–	*	–
59.21	6.77	5	*	*	*	*	–	–
58.79	7.16	5	*	*	*	–	–	*
58.72	7.22	5	–	*	*	*	–	*
57.51	8.34	5	–	*	*	–	*	*
55.87	9.86	5	*	*	–	*	*	–
55.52	10.19	5	–	*	*	*	*	–
52.74	12.76	5	*	*	–	*	–	*

Best subsets with 5 terms

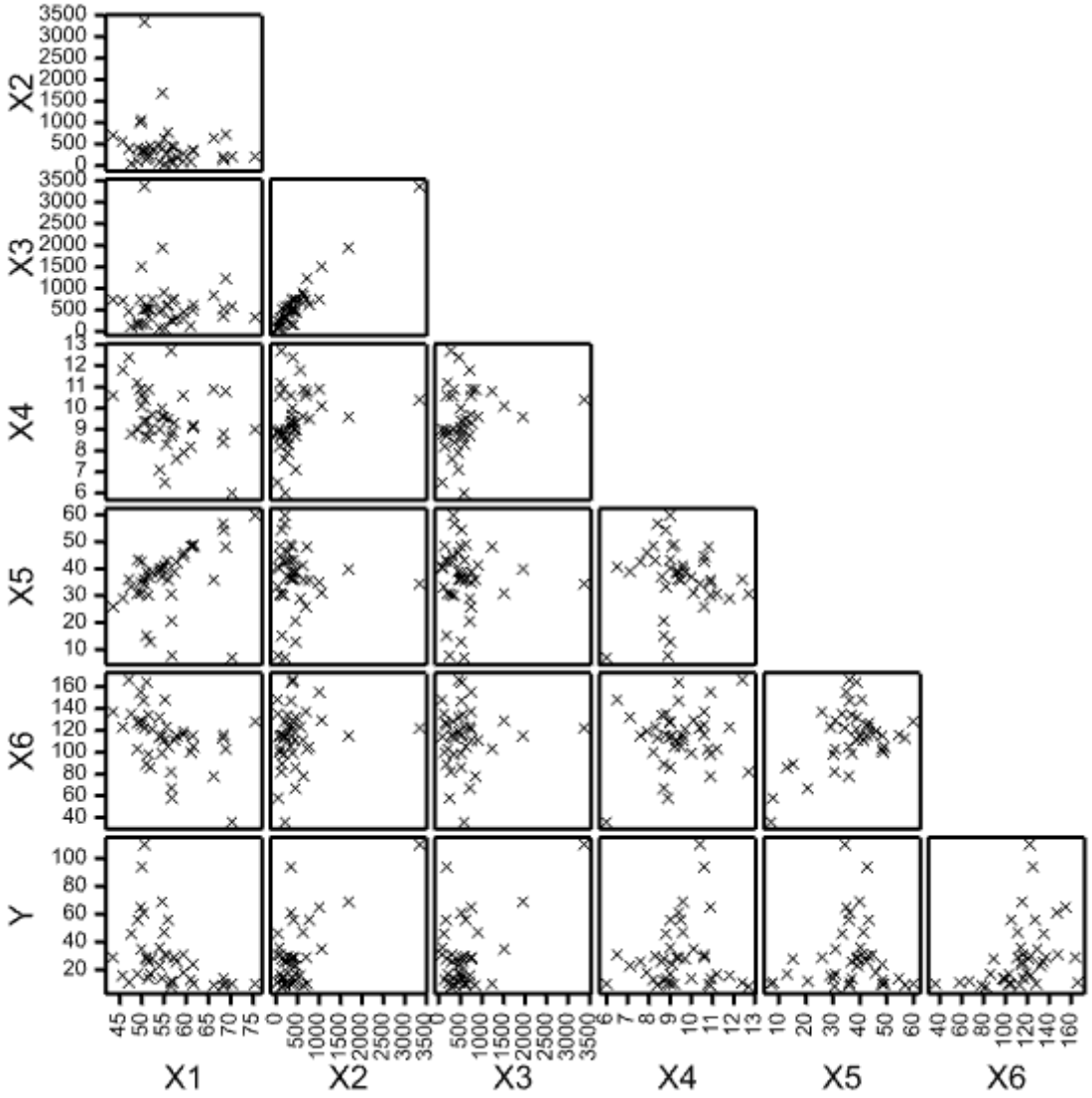
Adjusted R^2	C_p	Df	X1	X2	X3	X4	X5	X6
62.12	5.10	6	*	*	*	*	*	–
60.01	6.99	6	*	*	*	*	–	*
58.82	8.07	6	*	*	*	–	*	*
57.60	9.17	6	–	*	*	*	*	*
54.75	11.74	6	*	*	–	*	*	*
43.35	21.99	6	*	–	*	*	*	*

Best subsets with 6 terms

Adjusted R^2	C_p	Df	X1	X2	X3	X4	X5	X6
61.12	7.00	7	*	*	*	*	*	*

Estimates of parameters

Parameter	estimate	s.e.	t_{34} value	p-value
Constant	111.7	47.3	2.36	0.024
X1	-1.268	0.621	-2.04	0.049
X2	0.0649	0.0157	4.12	<.001
X3	-0.0393	0.0151	-2.60	0.014
X4	-3.18	1.82	-1.75	0.089
X5	0.512	0.363	1.41	0.167
X6	-0.052	0.162	-0.32	0.750



5, An agricultural experiment was conducted to investigate the response of two varieties of spring barley to levels of soil phosphorus, recorded on the Olsen P scale. The experiment was arranged as a randomised complete block design with 4 blocks (replicates), each containing 10 plots. Each of the two varieties of spring barley was grown at the same five levels of phosphorus, labelled as 2, 5, 8, 11 and 14 units.

The table below gives, in suitable units, the total yield for the four plots of each treatment combination, plus the totals for each of the two varieties, and for each of the five phosphorus levels.

Variety	*Phosphorus level (Olsen P)*					
	2	5	8	11	14	Totals
A	9.41	13.82	12.95	13.94	16.09	66.21
B	11.27	17.82	21.93	21.48	24.52	97.02
Totals	20.68	31.64	34.88	35.42	40.61	163.23

The four block totals (for 10 plots each) are 41.28, 42.06, 42.30 and 37.59, and the sum of squares for the 40 observations is 736.784. You are also given that the sum of squares for the 10 treatment totals ($9.41^2 + 13.82^2 + \ldots + 24.52^2$) is 2888.566.

(i) Construct an analysis of variance to assess the effects of variety, phosphorus and the interaction between these factors, taking full account of the design of the experiment, and comment on the results.

(8)

The experimenter is particularly interested in the shape of the response to the level of phosphorus.

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

(3)

(iii) Determine coefficients for orthogonal linear and quadratic polynomial contrasts with five equally-spaced levels, demonstrating that they are orthogonal using your answer to part (ii).

(3)

(iv) Calculate the contrast values and the associated sums of squares for the overall orthogonal linear and quadratic contrasts for the effect of phosphorus in the data above. Interpret the results in terms of the shape of the response to levels of phosphorus. 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 yields.

(6)

6. In an experiment assessing the impact of different concentrations of a chemical on flower production, three different species of flowering crop have been treated at the same five concentrations of the chemical (including a zero concentration to provide a baseline response). For each combination of species and concentration, the numbers of flowers produced have been recorded for eight replicate plants.

Summaries of the results of two analyses are provided **on the next two pages**, along with tables showing the means and variances of the observed counts for each treatment combination together with a plot of the variances against the means. For each analysis residual plots are also provided.

(i) Explain what is meant by a *variance-stabilising transformation* in the context of a linear model, and use the residual plots for the first analysis (of observed counts) to identify why a transformation might be needed before analysing these data.

(3)

(ii) A random variable Y has mean μ and standard deviation σ, and X is defined as a function $X = h(Y)$ of Y. Use a suitable Taylor series expansion to derive approximate expressions for $E(X)$ and $\text{Var}(X)$ in terms of the expected value and variance of Y. Hence show that if σ is a function $f(\mu)$ of the mean μ, an appropriate variance-stabilising transformation for Y might be $h(Y)$ as defined through the following relationship.

$$\frac{dh(Y)}{dY} = \frac{1}{f(Y)}$$

(4)

(iii) Use the result of part (ii) to suggest a suitable transformation for each of the following situations.

(a) The variance is proportional to the mean. (1)

(b) The standard deviation is proportional to the mean. (1)

(iv) Construct a plot of the standard deviations against the means for the observed counts, corresponding to the plot of the variances against the means provided on the next page. Using these plots, discuss which of the two transformations from part (iii) seems more appropriate for the analysis of these data. Comment further on whether the square root transformation appears appropriate by considering the residual plots provided on the second page following.

(4)

(v) Comparing the two presented analyses, comment on any additional benefits of applying the transformation.

(3)

(vi) Discuss alternative analysis approaches that might be applied, taking account of the likely underlying statistical distribution for these count data.

(4)

Output for Question 6 is on the next two pages

Analysis of variance – summary of the significance of model terms

p-value 1 is for the analysis of the observed counts, *p*-value 2 is for the analysis of the square-root transformed counts

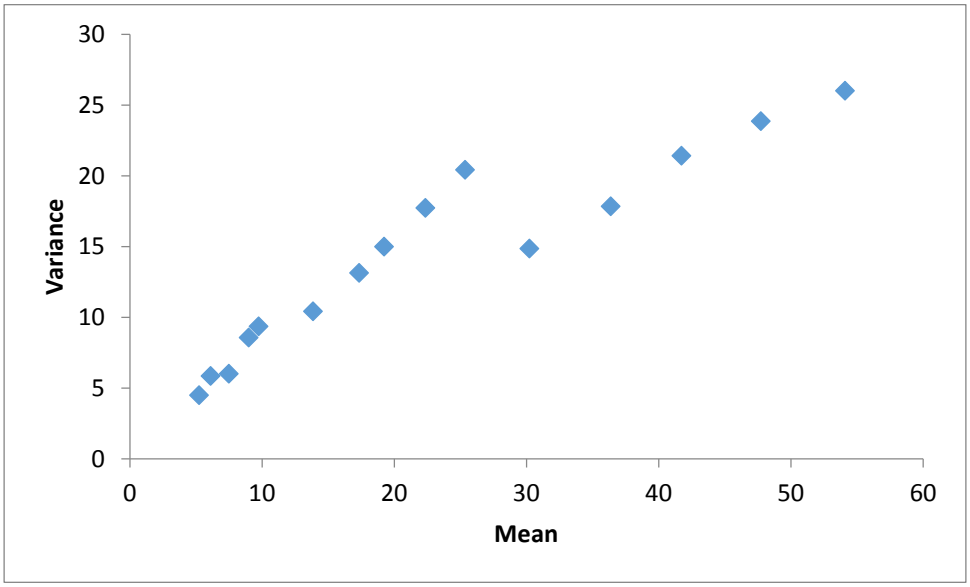
Source of variation	d.f.	*p*-value 1	*p*-value 2
Species	2	<.001	<0.001
Concentration	4	<.001	<0.001
Species.Concentration	8	0.031	0.694
Residual	105		
Total	119		

Means of observed counts

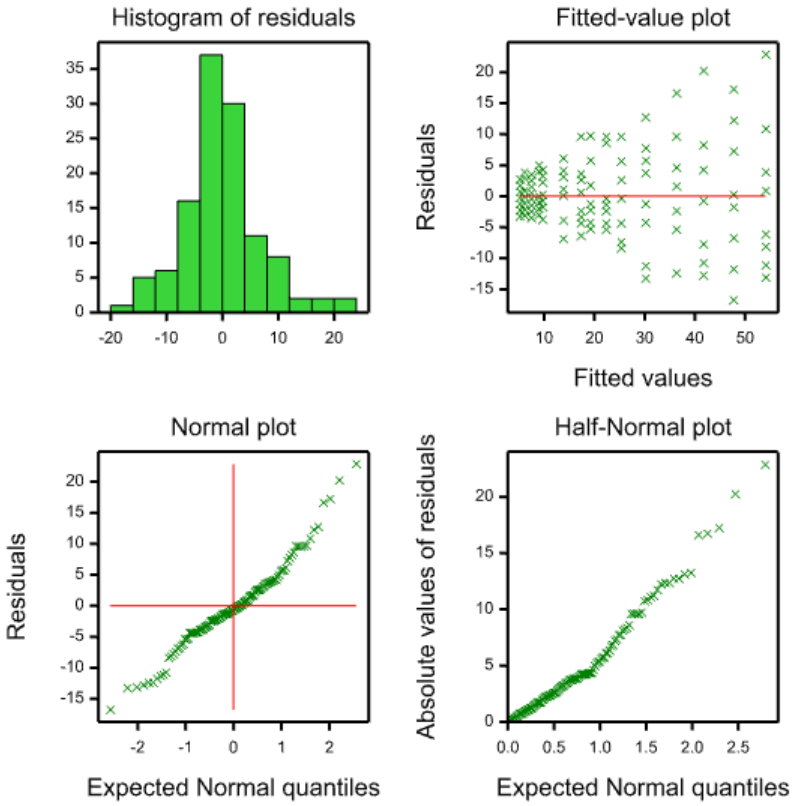
Species	Concentration				
	0	1	2	3	4
A	5.25	6.12	7.50	9.00	9.75
B	13.88	17.38	19.25	22.38	25.38
C	30.25	36.38	41.75	47.75	54.12

Variances of observed counts

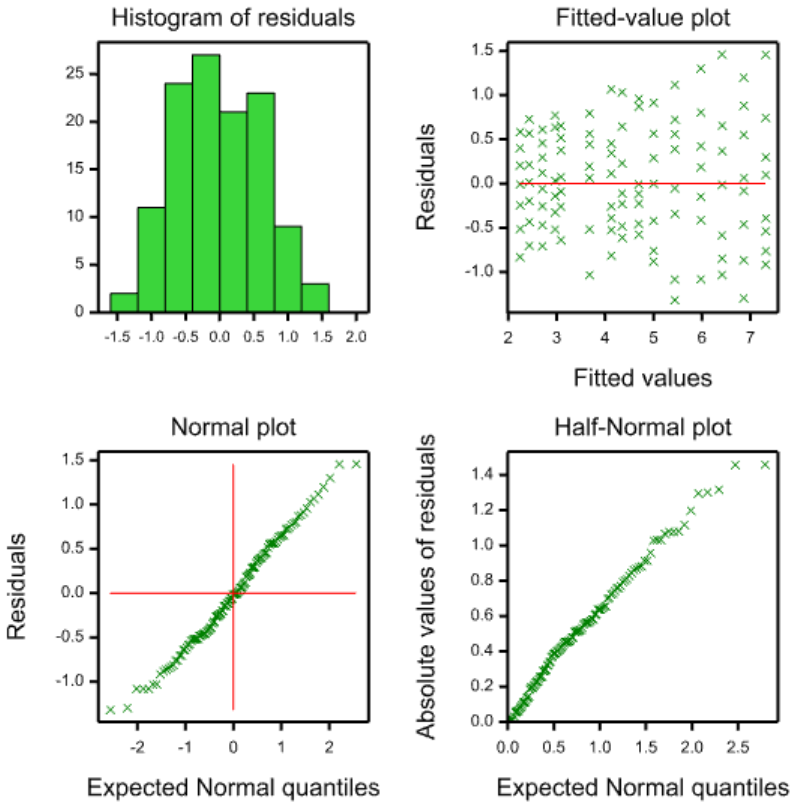
Species	Concentration				
	0	1	2	3	4
A	4.500	5.839	6.000	8.571	9.357
B	10.411	13.125	14.982	17.714	20.411
C	14.839	17.839	21.411	23.839	26.000



Residual plots for analysis of observed counts



Residual plots for analysis of square root transformed counts



7. (i) Identify the components of a generalised linear model (GLM), including an explanation of how the expected value of the response is related to the explanatory model.

(3)

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

(4)

An experiment was conducted to assess the efficacy of four new insecticides in controlling mosquitoes. Each insecticide was applied at seven concentrations (1ppm, 5ppm, 10ppm, 50ppm, 100ppm, 500ppm, 1000ppm), and each concentration of each insecticide was applied to six replicate batches of 10 mosquitoes each. The number of dead mosquitoes was recorded for each replicate of the 28 treatment combinations, and a generalised linear model analysis was applied with $\log_{10}$ concentration as the explanatory variable. The analysis assumed that the observed counts followed a binomial distribution and used a logit link function.

The results for three separate analyses are shown **on the next page** – the first analysis fitted a single line for the effect of insecticide concentration (i.e. ignoring any differences between the four insecticides), the second analysis fitted a single slope parameter but allowed for different intercept parameters for each insecticide, and the third analysis fitted separate slopes and intercepts for each of the different insecticides.

(iii) Interpret the results of the three presented analyses and identify the most appropriate model to describe the observed responses. Construct an accumulated analysis of deviance table to show the calculations and tests that you have used to reach your conclusion, and comment on any potential lack-of-fit.

(8)

(iv) Further output is provided for the second model **on the next page**, detailing the fitted parameter values and the estimated LD50s (the concentrations needed to kill 50% of the mosquitos). Explain how the LD50 values are obtained from the fitted parameters, and why, for this model, it is appropriate to calculate the relative efficacy (expressed in terms of the change in concentration required for a particular level of kill) of one insecticide to another. Calculate the relative efficacy of insecticide A to insecticide B.

(5)

Output for Question 7 is on the next page

Model 1 – Single line

Source	d.f.	deviance	mean deviance
Regression	1	916.3	916.30
Residual	166	244.4	1.47
Total	167	1160.7	6.95

Model 2 – Parallel lines (separate intercept parameter for each insecticide)

Source	d.f.	deviance	mean deviance
Regression	4	956.3	239.08
Residual	163	204.4	1.25
Total	167	1160.7	6.95

Model 3 – Separate lines (separate slope and intercept parameters for each insecticide)

Source	d.f.	deviance	mean deviance
Regression	7	957.0	136.71
Residual	160	203.7	1.27
Total	167	1160.7	6.95

Parameter estimates for Model 2

Parameter	estimate	s.e.
Insecticide A	-3.032	0.196
Insecticide B	-3.494	0.208
Insecticide C	-3.863	0.219
Insecticide D	-4.197	0.229
$\log_{10}$(Concentration)	2.1066	0.0955

LD50s for Model 2 (on $\log_{10}$ scale)

Insecticide	estimate	s.e.
A	1.440	0.06508
B	1.659	0.06484
C	1.834	0.06495
D	1.992	0.06534

8. (i) Explain how a generalised linear model (GLM) can be used to analyse contingency table data, describing the form of the link function and explanatory variables, and the distributional assumptions.

(3)

Following reports of a large number of respiratory infections by recent residents at a hotel holiday resort, public health officials conducted a survey of a large group of recent residents to identify possible causes. Data were collected about whether residents drank water from the taps in the hotel (Yes, No), whether they swam while at the hotel (in the Hotel Pool, on the Local Beach, or Not at all), which of the catering facilities they had used at the hotel (None, Restaurant only, Bar + Restaurant), and whether they had suffered with a respiratory infection. The numbers of residents for each combination of responses are shown below.

	Tap-water	Yes		No	
	Infection	Yes	No	Yes	No
Swimming	*Catering*				
Not at all	None	6	7	0	24
	Bar + Restaurant	13	4	6	12
	Restaurant only	5	4	3	14
Hotel Pool	None	12	8	7	7
	Bar + Restaurant	22	2	18	4
	Restaurant only	15	4	12	9
Local Beach	None	6	8	0	12
	Bar + Restaurant	12	5	4	13
	Restaurant only	6	4	2	6

A series of log-linear models, including different terms, has been fitted to these data to identify associations between possible causal factors and infection. The results are summarised in the table below, where RI = Respiratory Infection, TW = Tap-water, SW = Swimming and HC = Hotel Catering. In these models A*B is used as a shorthand to indicate that both of the main effects of A and B and the interaction between A and B are included, while A.B indicates the interaction between A and B.

Terms in model	*Residual df*	*Deviance*
RI + TW*SW*HC (Baseline)	17	112.80
Baseline + RI.TW	16	78.23
Baseline + RI.SW	15	74.67
Baseline + RI.HC	15	89.01
Baseline + RI.(TW + SW)	14	41.00
Baseline + RI.(TW + HC)	14	52.11
Baseline + RI.(SW + HC)	13	50.49
Baseline + RI.(TW*SW)	12	33.87
Baseline + RI.(TW*HC)	12	51.43
Baseline + RI.(SW*HC)	9	48.19
Baseline + RI.(TW + SW + HC)	12	14.28
Baseline + RI.(TW*SW + HC)	10	7.73
Baseline + RI.(TW*HC + SW)	10	13.99
Baseline + RI.(SW*HC + TW)	8	12.77
Baseline + RI.(TW*(SW + HC))	8	7.06
Baseline + RI.(SW*(TW + HC))	6	6.42
Baseline + RI.(HC*(TW + SW))	6	12.30
Baseline + RI.(TW*SW + TW*HC + SW*HC)	4	5.63

Question 8 continued on the next page

(ii) What would be the values of the deviance and the residual degrees of freedom for the saturated model that includes all the terms in the final model above plus the four-factor interaction? Describe the interpretation of this four-factor interaction, and identify whether there is any evidence for needing to use the saturated model to describe the observed data.

(3)

(iii) Explain what the terms included in the Baseline model represent, and therefore why we are only interested in considering models that are more complex than the Baseline model.

(4)

(iv) Using forward selection, identify the best model for the observed data in respect of fit and parsimony, showing all your reasoning at each step.

(6)

(v) Interpret this best model to identify the likely causes of the respiratory infections, explaining the model terms in language understandable to a non-statistician.

(4)

BLANK PAGE