ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

# EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

## GRADUATE DIPLOMA, 2014

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

There are 8 questions altogether in the paper.

1. In a study to explore the germination behaviour of the sclerotia (fungal resting bodies) that cause a particular plant disease, a pathologist has run an experiment to assess the impact of storage temperature (10°C, 15°C, 20°C) and storage duration (0, 1, 2, 3 or 4 weeks) on the subsequent germination of three different populations (labelled UK, France, Spain). For each combination of storage temperature (temperature), storage duration (time) and population (population), three replicate samples of 50 sclerotia were placed in Petri dishes, and the numbers of sclerotia that had germinated after 10 days under standard conditions were recorded. To analyse these data, the experimenter has first converted each count into a percentage germination value. The output from an analysis of variance of the percentage germination 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 the percentage germination changes in response to each of the terms included in the fitted model.

(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 sclerotia being tested, might cause one or more of these assumptions to be invalid. Identify how such a violation of 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 1 is on the next two pages**

**Analysis of variance**

| Source of variation | df | SS | MS | MS ratio | p-value |
|---|---|---|---|---|---|
| Population | 2 | 2228.33 | 1114.16 | 47.36 | <.001 |
| Temperature | 2 | 3361.48 | 1680.74 | 71.44 | <.001 |
| Time | 4 | 24566.99 | 6141.75 | 261.06 | <.001 |
| Population.Temperature | 4 | 155.14 | 38.79 | 1.65 | 0.169 |
| Population.Time | 8 | 1703.23 | 212.90 | 9.05 | <.001 |
| Temperature.Time | 8 | 1719.41 | 214.93 | 9.14 | <.001 |
| Population.Temperature.Time | 16 | 409.30 | 25.58 | 1.09 | 0.379 |
| Residual | 90 | 2117.33 | 23.53 | | |
| Total | 134 | 36261.21 | | | |

**Tables of means**

Grand mean 17.44

| Population | UK | France | Spain |
|---|---|---|---|
| | 12.22 | 17.96 | 22.13 |

| Temperature | 10 | 15 | 20 |
|---|---|---|---|
| | 11.29 | 17.51 | 23.51 |

| Time | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| | 1.63 | 5.78 | 15.56 | 25.41 | 38.81 |

| | | Temperature | | |
|---|---|---|---|---|
| | | 10 | 15 | 20 |
| Population | UK | 6.67 | 13.60 | 16.40 |
| | France | 10.93 | 17.20 | 25.73 |
| | Spain | 16.27 | 21.73 | 28.40 |

| | | Time | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| Population | UK | 1.33 | 4.00 | 11.33 | 18.67 | 25.78 |
| | France | 1.78 | 6.44 | 15.78 | 25.78 | 40.00 |
| | Spain | 1.78 | 6.89 | 19.56 | 31.78 | 50.67 |

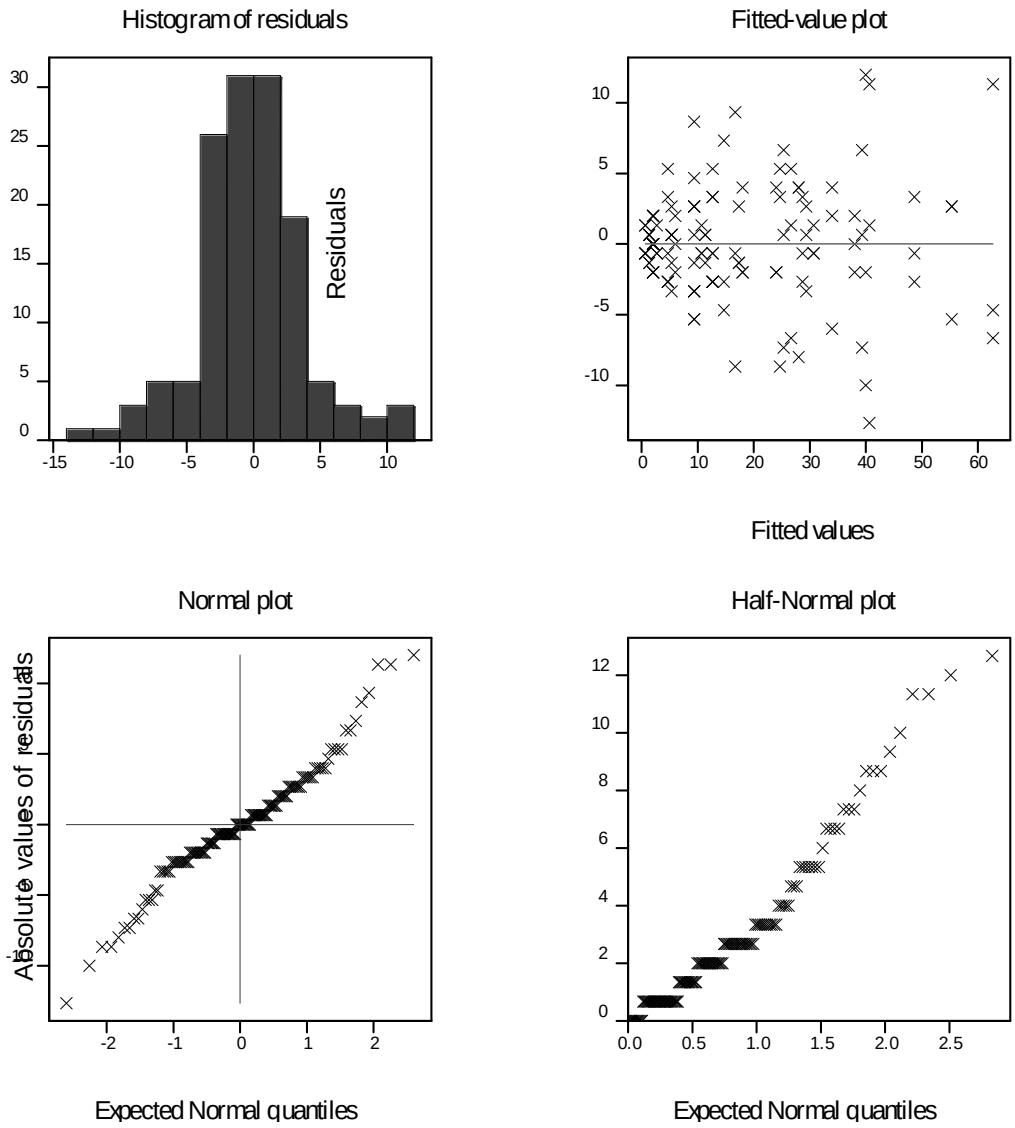
| | | Time | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| Temperature | 10 | 1.11 | 3.11 | 9.78 | 15.56 | 26.89 |
| | 15 | 1.78 | 6.22 | 15.78 | 24.89 | 38.89 |
| | 20 | 2.00 | 8.00 | 21.11 | 35.78 | 50.67 |

**Standard errors of differences of means**

| Table | Population | Temperature | Time | Population Temperature | Population Time | Temperature Time |
|---|---|---|---|---|---|---|
| rep. | 45 | 45 | 27 | 15 | 9 | 9 |
| s.e.d. | 1.023 | 1.023 | 1.320 | 1.771 | 2.286 | 2.286 |

**Output for Question 1 is continued on the next page**

**Residual plots for analysis of variance of percentage germination data**



2. An industrial experiment is to be performed to assess the impact of three production factors, A – C, and two environmental factors, D and E, each at two levels, on the quality of paint finish on car parts. Resources allow for a maximum of 48 processing runs to be included in the experiment, but the time taken per processing run means that no more than 8 runs can be included within each production shift. It is not expected that there will be strong interactions between the three production factors, but it is anticipated that there will be some interactions between the effects of the two environmental factors and between the environmental factors and the production factors.

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

(4)

(ii) Identify appropriate confounding and defining (fractionating) contrasts to generate a confounded half-replicate design for this experiment in two blocks of size 8, so that all main effects and appropriate two-factor interactions can be estimated. Identify the treatment combinations to be included in each block, and the aliasing structure.

(7)

(iii) Write down the outline of the analysis of variance table, listing the terms in it and their degrees of freedom. Briefly discuss why this confounded half-replicate design is unsatisfactory.

(4)

(iv) Describe how this design could be extended, using all 48 available processing runs, to provide satisfactory information about all main effects and the two-factor interactions of most interest. Briefly describe a further modification that could be made to allow the estimation of all two-factor interactions.

(5)

3. An experiment is to be performed to compare the responses of nine treatments A – I. There is no limitation on the number of replicates of each treatment that can be included in the experiment, but constraints on the number of experimental units that can be assessed on a single day mean that the maximum block size is seven experimental units.

(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 units per block, $k$.

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

Here, $r$ is the number of replicates of each treatment, $b$ is the number of blocks and $\lambda$ is the number of times that each pair of treatments appear together in a block.

(5)

(ii) Using the relationships in part (i), show that the smallest balanced incomplete block design for nine treatments with a block size of four units ($k = 4$) requires 72 units in total with eight replicates of each of the nine treatments ($r = 8$), 18 blocks ($b$) and with each pair of treatments occurring together in blocks three times ($\lambda = 3$).

(4)

(iii) Using the relationships in part (i), identify the smallest balanced incomplete block design for nine treatments with a block size of five experimental units, clearly identifying the values of the parameters $r$, $b$ and $\lambda$ in your answers together with the overall size of the experiment.

(3)

The smallest balanced incomplete block design for nine treatments in blocks of six experimental units requires 12 blocks, with each treatment being replicated eight times, and each pair of treatments occurring together in five different blocks. The allocation of treatments to blocks is obtained by first dividing the blocks into four sets of three blocks each, with each set of blocks containing two complete replicates of the nine treatments. The allocation is then obtained by considering the treatments in three groups of three treatments, {A, B, C}, {D, E, F} and {G, H, I}. In one set of blocks the treatments in two of the three groups are compared in each block, while in the other three sets of blocks, pairs of treatments from each group are compared in each block.

(iv) Use the information above to construct the treatment allocation for this balanced incomplete block design.

(5)

(v) Describe how this balanced incomplete block design should be randomised to be performed on 12 separate days, taking account of the order in which treatments are assessed during each day.

(3)

BLANK PAGE

4. An industrial experiment is concerned with the production of a complex polymer used in the plastic moulding industry. The experimenter is interested in enhancing the quantity of polymer produced using a standard method by the inclusion of two additional chemicals, each at two different concentrations. There are seven treatments as listed below. In addition three different machines are used for producing the polymer, and these may also have an impact on the level of production, an impact that may vary with the additional chemicals.

| *Code* | *Chemical Addition* |
|---|---|
| Standard | No additions |
| A1 | Chemical A at Low concentration |
| A2 | Chemical A at High concentration |
| B1 | Chemical B at Low concentration |
| B2 | Chemical B at High concentration |
| A1B1 | Chemical A at Low concentration /Chemical B at Low concentration |
| A2B2 | Chemical A at High concentration/Chemical B at High concentration |

Three replicate runs are completed for each of the 21 treatment combinations, with each replicate completed on a separate day. Within each day, the 21 runs are considered in a randomised order, so that the complete experiment can be considered to be arranged as a randomised complete block design with three blocks and 21 units per block.

The table below gives, in a suitable unit, the total production for the three plots for each treatment combination, plus the totals for each Machine and each Chemical Addition treatment.

| *Chemical Addition* | *Machine* | | | |
|---|---|---|---|---|
| | M1 | M2 | M3 | Totals |
| Standard | 173 | 160 | 225 | 558 |
| A1 | 308 | 186 | 373 | 867 |
| A2 | 388 | 291 | 420 | 1099 |
| B1 | 256 | 227 | 294 | 777 |
| B2 | 371 | 297 | 421 | 1089 |
| A1B1 | 358 | 293 | 367 | 1018 |
| A2B2 | 417 | 357 | 490 | 1264 |
| Totals | 2271 | 1811 | 2590 | 6672 |

The three block totals (for 21 runs each) are 2343, 2077 and 2252, and the sum of squares for the 63 observations is 766 924. You may also use the fact that the sum of squares for the 21 treatment totals ($173^2 + 160^2 + \ldots + 490^2$) is 2 283 100.

(i) Construct an analysis of variance to assess the effects of Machine, Chemical Addition and the interaction between these factors, and comment on the results. (7)

**Question 4 is continued on the next page**

Comparisons among the seven Chemical Addition treatments could be further explored using a set of six orthogonal contrasts.

(ii) Explain why the use of a set of orthogonal contrasts is particularly helpful in exploring comparisons such as these, and identify how the coefficients for any two contrasts can be used to check that they are orthogonal.

(3)

(iii) Identify six orthogonal contrasts appropriate for exploring the comparisons among the seven Chemical Addition treatments, including a comparison of the High and Low concentrations, a comparison of adding one chemical (A or B) and adding both chemicals, and a comparison between adding chemical A and adding chemical B, indicating the coefficients for each treatment in each contrast.

(6)

(iv) Calculate the contrast and associated sum of squares for a comparison of adding chemical A and adding chemical B, and for a comparison of the High and Low concentrations, and interpret the results in terms of the importance of these comparisons.

(4)

5. (i) Identify the components of a generalised linear model (GLM), using an equation to show the relationship between them. (4)

(ii) Specify these components for the log-linear model that underlies analysis of data in the form of a contingency table. (3)

Following the conference dinner at a major statistical conference, a number of the conference delegates were diagnosed as suffering from food poisoning. Data were available on the different menu choices that each delegate had made for each of three courses, and these data were analysed in an attempt to identify the likely cause(s) of the food poisoning. There were three choices for the starter (Soup, Prawn Cocktail, Paté), three for the main course (Chicken, Beef, Nut Roast) and two for the sweet (Fruit Salad, Chocolate Gateaux). The numbers of delegates selecting each combination, and the numbers of cases with and without food poisoning for each combination, are shown below.

| | *Sweet* | Fruit Salad | | Chocolate Gateaux | |
|---|---|---|---|---|---|
| | *Food Poisoning* | Yes | No | Yes | No |
| *Starter* | *Main* | | | | |
| Soup | Chicken | 12 | 3 | 15 | 7 |
| | Beef | 5 | 6 | 1 | 9 |
| | Nut Roast | 4 | 3 | 2 | 7 |
| Prawn Cocktail | Chicken | 11 | 2 | 14 | 8 |
| | Beef | 7 | 4 | 1 | 11 |
| | Nut Roast | 4 | 1 | 0 | 4 |
| Paté | Chicken | 13 | 5 | 16 | 8 |
| | Beef | 7 | 2 | 1 | 10 |
| | Nut Roast | 7 | 5 | 0 | 6 |

A series of log-linear models has been fitted to attempt to identify the likely cause(s) of the food poisoning, including different terms (where FP = Food Poisoning, ST = Starter, M = Main and SW = Sweet). The results are summarised in the table below. 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* |
|---|---|---|
| FP + ST*M*SW (Baseline) | 17 | 70.77 |
| Baseline + FP.ST | 15 | 70.66 |
| Baseline + FP.M | 15 | 43.30 |
| Baseline + FP.SW | 16 | 53.60 |
| Baseline + FP.(ST + M) | 13 | 43.23 |
| Baseline + FP.(ST + SW) | 14 | 53.47 |
| Baseline + FP.(M + SW) | 14 | 16.80 |
| Baseline + FP.(ST*M) | 9 | 42.43 |
| Baseline + FP.(ST*SW) | 12 | 52.35 |
| Baseline + FP.(M*SW) | 12 | 7.26 |
| Baseline + FP.(ST + M + SW) | 12 | 16.68 |
| Baseline + FP.(ST*M + SW) | 8 | 15.17 |
| Baseline + FP.(ST*SW + M) | 10 | 15.06 |
| Baseline + FP.(M*SW + ST) | 10 | 7.16 |
| Baseline + FP.(ST*(M + SW)) | 6 | 13.67 |
| Baseline + FP.(M*(ST + SW)) | 6 | 4.79 |
| Baseline + FP.(SW*(ST + M)) | 8 | 5.48 |
| Baseline + FP.(ST*M + ST*SW + M*SW) | 4 | 3.39 |

**Question 5 is continued on the next page**

(iii) What would be the values of the deviance and the residual degrees of freedom for the saturated model which includes all the terms in the final model above plus the four-factor interaction?

(1)

(iv) 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 this Baseline model.

(3)

(v) Using backward elimination, identify the best model for the data in terms of fit and parsimony, showing all your reasoning at each step.

(6)

(vi) Interpret this best model to identify the likely cause(s) of the food poisoning, explaining the model terms in language understandable to a non-statistician.

(3)

6. (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.

(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)

A weed scientist has harvested seed for 30 different species of grasses and taken various different measurements on samples of seed for each species – weight for 1000 seeds (We), mean length (L), mean width (Wi), mean hardness (Ha), harvest date (days after 31 December) (HD), and median number of seeds per seed head (N). She has also calculated a measure of the shape of the seeds (by dividing the mean length by the mean width) (S) and of the surface area (by multiplying the mean length by the mean width) (A). She has then used a multiple linear regression analysis to determine whether there is a relationship between 1000-seed weight (We) and some combination of the other measurements made.

The analysis output shows the correlation matrix for the eight measured and calculated variables, the summary of a forward selection variable selection process where terms are included in the model if the variance ratio is greater than 2.00, and the fitted parameters for the model finally selected using the forward selection process.

**Correlation matrix**

```
L    0.4158
Wi   0.8619  -0.0368
Ha   0.0808  -0.1156   0.1113
HD  -0.1206  -0.1168   0.0189  -0.1615
N   -0.2698   0.0644  -0.3991  -0.1219   0.3220
S   -0.4009   0.5783  -0.7333  -0.2287  -0.0271   0.2388
A    0.9662   0.5991   0.7552   0.0571  -0.1017  -0.2034  -0.2633
        We       L        Wi       Ha       HD       N        S
```

**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:   460.8 - Adding Area              1782.9 - Adding Width
         5736.9 - Adding Length            5821.5 - Adding Shape
         6431.1 - Adding Number            6697.0 - No change
         6835.3 - Adding Harvest Date      6890.9 - Adding Hardness
Chosen action: Adding Area

Step 2:   179.7 - Adding Length              185.0 - Adding Width
          312.0 - Adding Shape               437.6 - Adding Number
          460.8 - No change                  473.1 - Adding Hardness
          474.3 - Adding Harvest Date      6697.0 - Dropping Area
Chosen action: Adding Length
```

**Question 6 is continued on the next page**

```
Step 3:   42.73 - Adding Shape            176.54 - Adding Harvest Date
         178.41 - Adding Width            179.73 - No change
         181.08 - Adding Number           185.44 - Adding Hardness
         460.83 - Dropping Length        5736.94 - Dropping Area
Chosen action: Adding Shape

Step 4:   30.97 - Adding Harvest Date      40.37 - Adding Width
          42.73 - No change                43.11 - Adding Number
          44.00 - Adding Hardness         179.73 - Dropping Shape
         312.01 - Dropping Length        1504.24 - Dropping Area
Chosen action: Adding Harvest Date

Step 5:   30.74 - Adding Width             30.97 - No change
          32.24 - Adding Number            32.25 - Adding Hardness
          42.73 - Dropping Harvest Date   176.54 - Dropping Shape
         316.83 - Dropping Length        1553.75 - Dropping Area
Chosen action: No change
```

**Accumulated analysis of variance**

| Change | df | SS | MS | MS ratio | p-value |
|---|---|---|---|---|---|
| + Area | 1 | 181310.23 | 181310.23 | 5855.27 | <.001 |
| + Length | 1 | 8050.41 | 8050.41 | 259.98 | <.001 |
| + Shape | 1 | 3741.70 | 3741.70 | 120.84 | <.001 |
| + Harvest Date | 1 | 336.97 | 336.97 | 10.88 | 0.003 |
| Residual | 25 | 774.13 | 30.97 | | |
| Total | 29 | 194213.44 | 6697.02 | | |

Final model: Constant + Area + Length + Shape + Harvest Date

**Parameters for chosen model**

| Parameter | estimate | s.e. | t value | p-value |
|---|---|---|---|---|
| Constant | 52.0 | 16.0 | 3.26 | 0.003 |
| Area | 2.2411 | 0.0627 | 35.77 | <.001 |
| Length | -9.067 | 0.584 | -15.52 | <.001 |
| Shape | 10.924 | 0.984 | 11.10 | <.001 |
| Harvest Date | -0.1950 | 0.0591 | -3.30 | 0.003 |

(iii) Describe how choice of the threshold value for the inclusion of terms in a forward selection stepwise process will influence the number of terms added to the model. Identify why the chosen action in Step 5 is "No change" even though "Adding Width" results in the smallest residual mean square.

(4)

(iv) Interpret the results from the forward selection stepwise process, including an explanation of why the Shape predictor variable has a positive coefficient (estimate) when the correlation between Weight and Shape is negative.

(5)

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

(4)

7. (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)

An experimental study was concerned with the growth patterns of colonies for a fungal pathogen under constant environmental conditions. Fungal colonies were grown on standard media in Petri dishes for four different isolates of the pathogen (A – D). The fungal colonies were destructively harvested, with three replicate dishes being harvested each day for 10 days after the start of the experiment, and the weight of each fungal colony measured. It was anticipated that the growth patterns would follow a sigmoidal growth curve, and so a logistic function

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

was fitted to the data for each isolate. The observed data and fitted growth curves for each of the four isolates are shown in the graphs **on the next page**, and the fitted parameter values (with standard errors) and ANOVA summary statistics for each isolate are shown in the table below.

| *Isolate* | A | B | C | D |
|---|---|---|---|---|
| Parameter *a* | –1.41 (2.26) | –0.85 (1.54) | –0.79 (1.56) | –0.76 (2.06) |
| Parameter *b* | 0.561 (0.173) | 0.890 (0.171) | 0.531 (0.127) | 0.480 (0.193) |
| Parameter *c* | 20.84 (3.81) | 20.58 (1.89) | 23.43 (3.53) | 25.23 (7.88) |
| Parameter *m* | 5.240 (0.426) | 4.037 (0.244) | 6.067 (0.359) | 7.207 (0.969) |
| *F*-value | 132.31 | 206.43 | 243.38 | 111.95 |
| *p*-value | <0.001 | <0.001 | <0.001 | <0.001 |
| Adjusted $R^2$ (%) | 93.1 | 95.5 | 96.2 | 92.0 |

In a further analysis of the complete dataset, a series of models was fitted to assess whether any of the remaining parameters could be assumed to be the same for the different isolates. A summary of the residual degrees of freedom and residual sums of squares for each of these models is shown below.

| *Model* | *Res df* | *Res SS* | *Model* | *Res df* | *Res SS* |
|---|---|---|---|---|---|
| Single line | 116 | 697.61 | Parameters *b*, *c* vary | 110 | 429.34 |
| Parameter *a* varies | 113 | 436.30 | Parameters *b*, *m* vary | 110 | 305.77 |
| Parameter *b* varies | 113 | 459.61 | Parameters *c*, *m* vary | 110 | 322.37 |
| Parameter *c* varies | 113 | 447.61 | Parameters *a*, *b*, *c* vary | 107 | * |
| Parameter *m* varies | 113 | 696.20 | Parameters *a*, *b*, *m* vary | 107 | 291.91 |
| Parameters *a*, *b* vary | 110 | 422.39 | Parameters *a*, *c*, *m* vary | 107 | 297.75 |
| Parameters *a*, *c* vary | 110 | 425.85 | Parameters *b*, *c*, *m* vary | 107 | 289.73 |
| Parameters *a*, *m* vary | 110 | 326.88 | Parameters *a*, *b*, *c*, *m* vary | 104 | 289.54 |

*Model fitting failed to converge

**Question 7 is continued on the next page**

(ii) Describe how the parameters of the logistic function relate to the sigmoidal shape of the fitted curve. Based on the fitted curves and parameter estimates, identify whether this function is appropriate for the observed data, and, if not, identify possible modifications or alternative models.

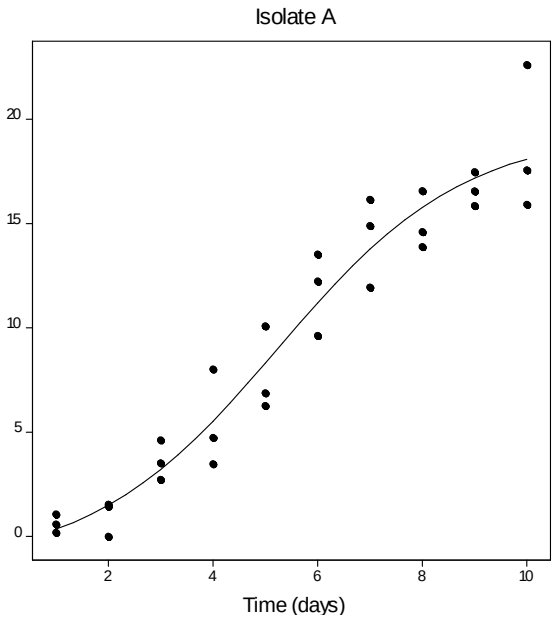
(5)

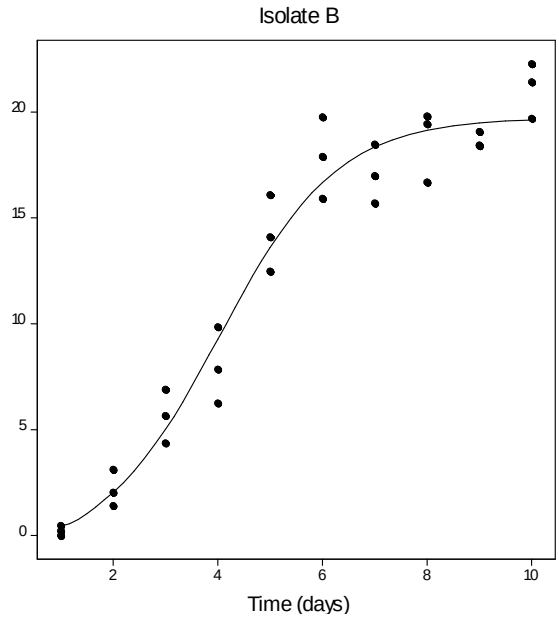
(iii) Describe how the analysis for each isolate could be extended to provide a test of the lack-of-fit of the model relative to the between-replicate variation.

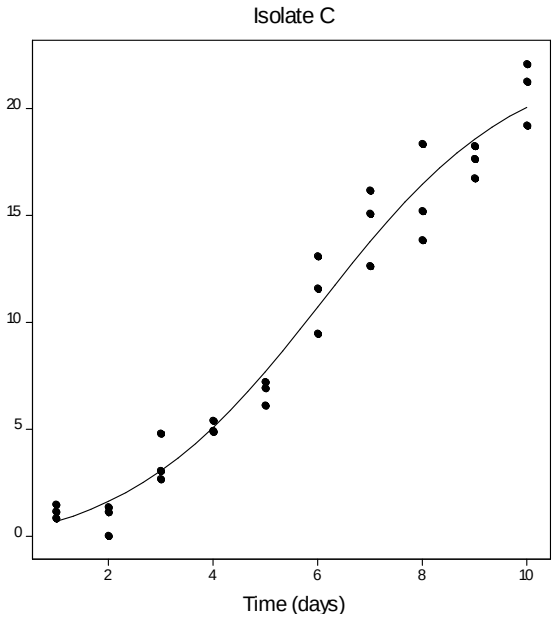
(3)

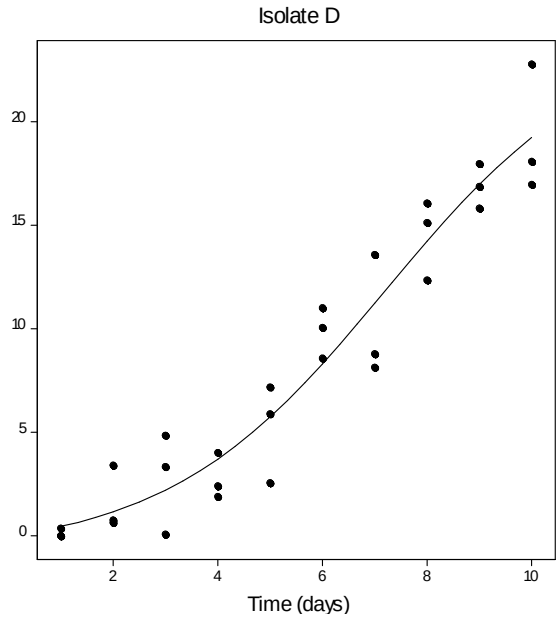
(iv) Use the summary of residual sums of squares for the series of fitted models to determine the best model to describe the complete dataset across all four isolates, and interpret this in terms of the fitted parameter values shown for the models fitted to individual isolates.

(6)



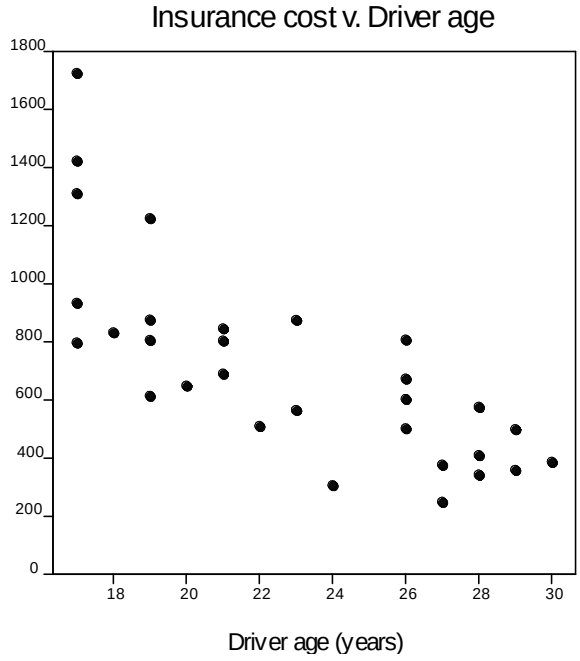
8. (i) Explain why it might be necessary to transform a response variable in a linear model to stabilise the variance, and give an example of where such a transformation might be useful in linear regression.

(3)

(ii) Explain how a weighted least squares approach is different from an ordinary least squares approach for linear regression, and state the conditions under which you might consider using a weighted least squares approach.

(3)

Data collected in a survey of the readers of a motoring magazine were used to assess the form of the relationship between the cost of motoring insurance and the age of the driver, for relatively young drivers. A scatter plot showing a representative subset of the collected data is shown below.




(iii) Describe any apparent relationship between the cost of insurance and driver age, as shown in the scatter plot, and comment on whether a transformation of the response variable or use of a weighted least squares analysis would be appropriate.

(6)

The output shown **on the next two pages** summarises the results from three possible analyses relating insurance cost to driver age:

1. a simple linear regression of the untransformed insurance costs;
2. a weighted linear regression of the untransformed insurance costs (using the reciprocal of the square of the insurance cost as the weight);
3. a simple linear regression of the log-transformed insurance costs.

**Question 8 is continued on the next page**

(iv) Interpret the output, including the diagnostic plots, identifying, with justification, the model that you consider best describes the response, and stating the fitted relationship for this best-fitting model. Your interpretation should include consideration of how useful the chosen model might be over a wider age range (i.e. including data for older drivers).

(8)

**Analysis 1**

**Summary of analysis**

| Source | d.f. | s.s. | m.s. | v.r. |
|---|---|---|---|---|
| Regression | 1 | 1979488 | 1979488 | 37.24 |
| Residual | 28 | 1488181 | 53149 | |
| Total | 29 | 3467669 | 119575 | |

Percentage variance accounted for 55.6
Standard error of observations is estimated to be 231.

**Estimates of parameters**

| Parameter | estimate | s.e. | t value |
|---|---|---|---|
| Constant | 2082 | 227 | 9.17 |
| age | -59.73 | 9.79 | -6.10 |

**Analysis 2**

**Summary of analysis**

| Source | d.f. | s.s. | m.s. | v.r. |
|---|---|---|---|---|
| Regression | 1 | 2.616 | 2.61572 | 26.50 |
| Residual | 28 | 2.764 | 0.09870 | |
| Total | 29 | 5.379 | 0.18549 | |

Percentage variance accounted for 46.8
Standard error of observations is estimated to be 0.314

**Estimates of parameters**

| Parameter | estimate | s.e. | t value |
|---|---|---|---|
| Constant | 1598 | 219 | 7.28 |
| age | -44.00 | 8.55 | -5.15 |

**Analysis 3**

**Summary of analysis**

| Source | d.f. | s.s. | m.s. | v.r. |
|---|---|---|---|---|
| Regression | 1 | 0.7330 | 0.73300 | 44.56 |
| Residual | 28 | 0.4606 | 0.01645 | |
| Total | 29 | 1.1936 | 0.04116 | |

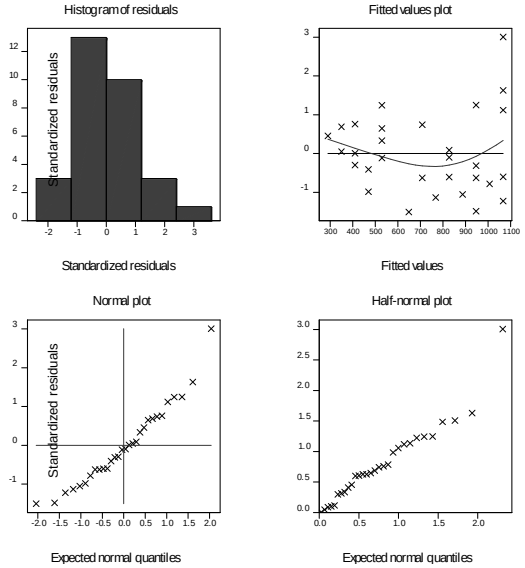
Percentage variance accounted for 60.0
Standard error of observations is estimated to be 0.128

**Estimates of parameters**

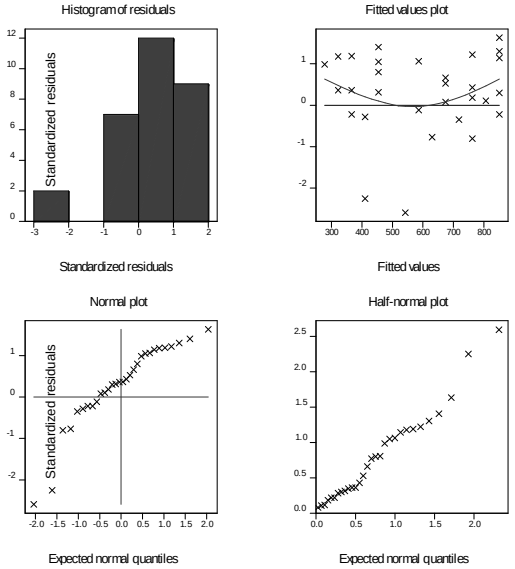
| Parameter | estimate | s.e. | t value |
|---|---|---|---|
| Constant | 3.641 | 0.126 | 28.82 |
| age | -0.03635 | 0.00545 | -6.68 |

**Output for Question 8 is continued on the next page**

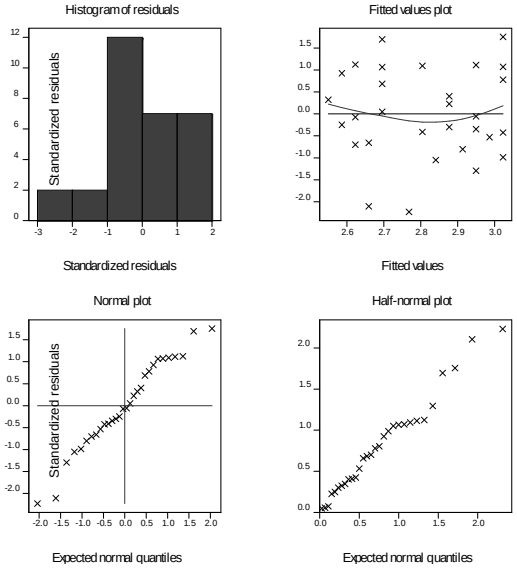
**Analysis 1 – diagnostic plots**




**Analysis 2 – diagnostic plots**




**Analysis 3 – diagnostic plots**



BLANK PAGE

BLANK PAGE