ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2017

MODULE 4 : Modelling experimental data

Time allowed: Three hours

Candidates should answer **FIVE** *questions.*

All questions carry equal marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

The notation log *denotes logarithm to base* ***e****.*
Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Note also that $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$.

This examination paper consists of 16 printed pages.
This front cover is page 1.
Question 1 starts on page 3.

There are 8 questions altogether in the paper.

BLANK PAGE

1. An experiment is to be conducted to compare the responses of seven treatments A – G. There is sufficient material available for 12 replicate samples to be processed using each of the seven treatments, and up to eight 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 the initial step can only produce enough material for a maximum of five samples, batches of samples from this initial step can be stored for processing at a later date.

(i) Identify the advantages and disadvantages associated with using a randomised complete block design rather than 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, 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.

(3)

(iii) One of the possible balanced incomplete block designs that could be constructed to cope with the constraints described above for this experiment has the following parameters as defined in part (ii): $t = 7$, $r = 8$, $k = 4$, $b = 14$, $\lambda = 4$. Find the parameter values for three other possible such designs. Discuss which of these four designs would make the most effective use of the available resources.

(6)

(iv) Describe how you would construct the allocation of treatment to blocks for the balanced incomplete block design with parameters $t = 7$, $r = 8$, $k = 4$, $b = 14$, $\lambda = 4$, and write down the allocation of treatments to blocks for this balanced incomplete block design.

(7)

2. (i) Identify the components of a generalised linear model (GLM), showing how the expected value of the response is related to the explanatory model.

(4)

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

(3)

In an ecological study, the presence or absence of an important plant species (used as an indicator of soil health) was recorded in a random selection of woodland sites across the UK, together with three categorical environmental variables – soil type (clay loam, silt loam, sandy loam), annual rainfall (low, medium, high), and altitude (low, high). The numbers of sites for each combination of the three environmental variables, and the numbers of sites with and without the indicator plant species for each combination, are shown below.

	Altitude	Low		High	
	Indicator Species	Present	Absent	Present	Absent
Soil Type	Rainfall				
Clay Loam	Low	12	4	5	10
	Medium	12	6	7	9
	High	11	5	6	11
Silt Loam	Low	10	7	8	11
	Medium	11	9	6	8
	High	7	5	9	13
Sandy Loam	Low	6	12	10	8
	Medium	4	13	11	7
	High	2	19	14	3

A series of log-linear models has been fitted in an attempt to identify how the environmental variables influence the health of the woodland soils, including different terms (where SH = Soil Health indicator species, ST = Soil Type, AR = annual rainfall and AL = altitude). The results are summarised in the table **on the next page**. In these models X*Y is used as a shorthand to indicate that both the main effects of X and Y and the interaction between X and Y are included, while X.Y indicates the interaction between X and Y.

Question 2 continues on the next page

Terms in model	*Residual df*	*Deviance*
SH + ST*AR*AL (Baseline)	17	46.07
Baseline + SH.ST	15	43.57
Baseline + SH.AR	15	45.85
Baseline + SH.AL	16	46.07
Baseline + SH.(ST+AR)	13	43.36
Baseline + SH.(ST+AL)	14	43.57
Baseline + SH.(AR+AL)	14	45.84
Baseline + SH.(ST*AR)	9	43.31
Baseline + SH.(ST*AL)	12	7.60
Baseline + SH.(AR*AL)	12	43.40
Baseline + SH.(ST+AR+AL)	12	43.36
Baseline + SH.(ST*AR+AL)	8	43.30
Baseline + SH.(ST*AL+AR)	10	7.56
Baseline + SH.(AR*AL+ST)	10	41.03
Baseline + SH.(ST*(AR+AL))	6	7.45
Baseline + SH.(AR*(ST+AL))	6	40.94
Baseline + SH.(AL*(ST+AR))	8	5.03
Baseline + SH.(ST*AR+ST*AL+AR*AL)	4	4.90

(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 the fit and parsimony, showing all your reasoning at each step.

(6)

(vi) Interpret this best model to identify the key environmental variables influencing soil health, explaining the model terms in language understandable to a non-statistician.

(3)

3. (i) Discuss the differences between linear and non-linear regression models, identifying examples of each, 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 into a linear form.

(a) $y = ae^{bx}$ [parameters a, b]

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

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

(d) $y = \dfrac{ax}{1 + bx}$ [parameters a, b]

(4)

An experiment was performed to assess the effect of nitrogen fertiliser on the production of sugar cane. The data were obtained from 4 replicate plots of each of 5 different levels of nitrogen fertiliser (0, 50, 100, 150, 200), with the experiment arranged following a completely randomised design. An exponential model (with parameters A, B, and R, and with X representing the explanatory variable (nitrogen fertiliser rate)) was fitted to describe the observed response, and the results from this model fitting are shown **on the next page**, including a plot of the fitted model and observed data. (Here R**X denotes R^X.)

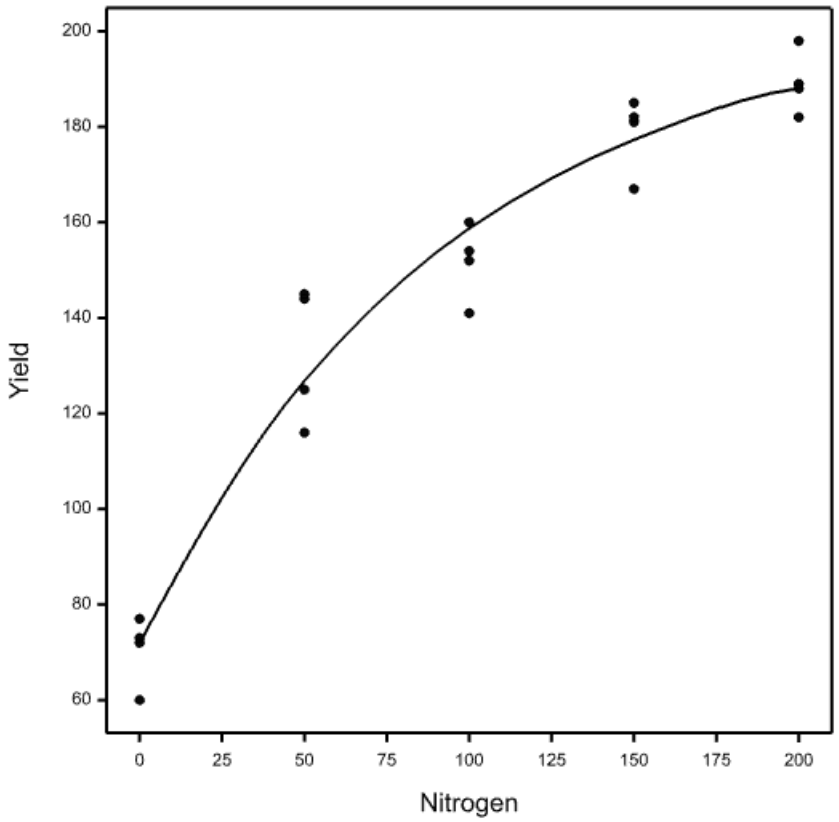
(iii) Interpret the results of the analysis, including a description of how the fitted parameters relate to the sugar yield response to nitrogen. Estimate the level of nitrogen needed to achieve a sugar yield of 170.

(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 3 is on the next page



Nonlinear regression analysis

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

Summary of analysis

Source	df	SS	MS	MS ratio
Regression	2	35046	17523.18	182.02
Residual	17	1637	96.27	
Total	19	36683	1930.68	

Percentage variance accounted for 95.0
Standard error of observations is estimated to be 9.81.

Estimates of parameters

Parameter	estimate	s.e.
R	0.98920	0.00213
B	-131.1	10.6
A	203.0	10.8

4. A glasshouse trial is planned to assess the effect of two fungicide sprays on controlling a plant disease affecting the growth of wheat plants, in comparison with a "no fungicide" control treatment. Three different varieties, having different levels of genetic resistance to the plant disease, are to be included, with all nine combinations of the three varieties and three fungicide treatments to be included in the trial.

The experimental unit will be a pot in which a fixed number of wheat seeds are sown, with the plants inoculated with the plant disease pathogen at a defined growth stage. The fungicide treatments will then be applied to the plants in each pot at a fixed time after inoculation.

The glasshouse has space for up to 81 pots arranged in an array of 9 rows and 9 columns, the columns running from North to South and the rows from East to West. It is anticipated that there will be a temperature gradient from North to South, with variation in natural lighting levels from East to West.

(i) Identify the advantages and disadvantages of arranging the 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 pots would be arranged in the glasshouse, and show a dummy analysis of variance table.

(12)

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

(4)

(iii) The application of the fungicide sprays is most effective when applied to larger groups of plants, so the experimenter proposes to arrange the pots in groups to which the different spray treatments would be applied, with pots for each of the different varieties included in each group. Describe a suitable design approach for this scenario (including a sketch), and describe the randomisation process that should be followed for the allocation of treatments following this design.

(4)

5. An industrial experiment has been performed to assess the impact of the operating temperature on the production of a chemical. Primary interest is in responses at temperatures between 5 °C and 30 °C, but the experimenter is also interested in how the process works at higher temperatures up to 80 °C. There are three different sources of raw materials (labelled A, B and C), and two additives (labelled Y and Z) that are believed to enhance the production process. All combinations of raw material sources and additives have been included in the experiment, together with each of the raw materials in the absence of any additive (labelled X), and three replicate reactions have been observed at each of 8 temperatures for the 9 combinations of raw material source and additive.

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

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

(4)

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

(5)

(iii) Given the set of treatments considered in this experiment, briefly describe how the analysis might be extended to provide more detailed information about the impacts of the different raw material sources and additives of chemical production.

(3)

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

(4)

(v) The analyses of models 2 and 4 both identify 27 points with large (but equal) values of leverage. Describe the concepts of *leverage* and *influence* in regression analysis for the usual general linear model $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, and state how leverage values are obtained from the "hat" matrix $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. Identify the set of observations with the large leverage values.

(4)

Output for Question 5 is on the next two pages

Accumulated Analysis of Variance Table

```
Model 1:  prod ~ 1 (null model)
Model 2:  prod ~ temp
Model 3:  prod ~ temp + treat
Model 4:  prod ~ temp * treat
```

Model	Residual df	RSS	df	SS	MS ratio	p-value
1	215	1615380				
2	214	206085	1	1409296	3642.85	<0.001
3	206	80334	8	125751	40.63	<0.001
4	198	76600	8	3734	1.21	0.297

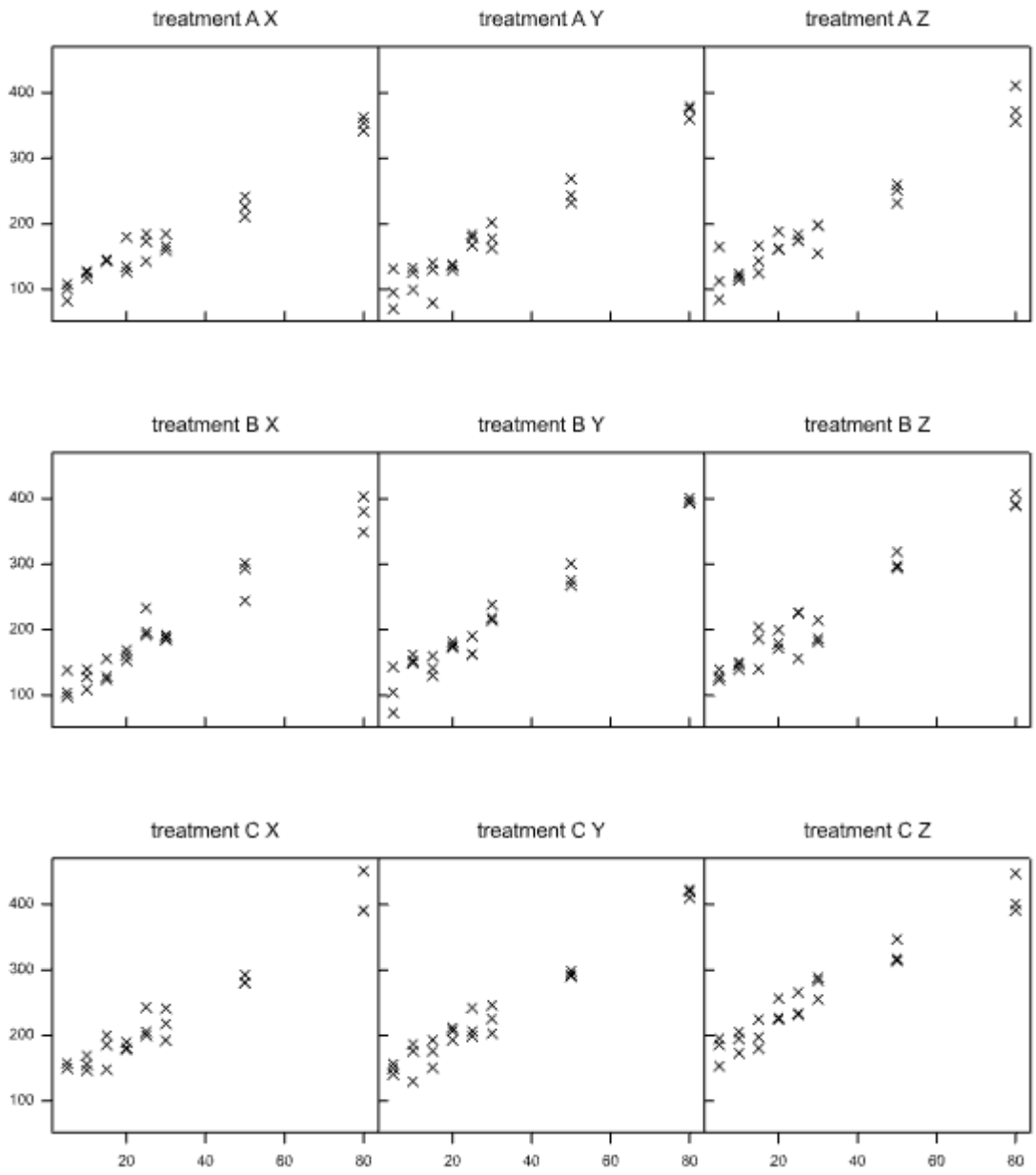
Model 2: prod ~ temp

Parameter	estimate	s.e.	t value	p-value
Constant	107.32	3.42	31.43	<.001
temp	3.4955	0.0914	38.25	<.001

Model 3: prod ~ temp + treat

Parameter	estimate	s.e.	t value	p-value
Constant	76.09	4.38	17.38	<.001
temp	3.4955	0.0581	60.12	<.001
treatment A Y	2.39	5.70	0.42	0.676
treatment A Z	14.64	5.70	2.57	0.011
treatment B X	20.00	5.70	3.51	<.001
treatment B Y	28.37	5.70	4.98	<.001
treatment B Z	38.02	5.70	6.67	<.001
treatment C X	46.64	5.70	8.18	<.001
treatment C Y	51.58	5.70	9.05	<.001
treatment C Z	79.45	5.70	13.94	<.001

Output for Question 5 continues on the next page


treatment A X
treatment A Y
treatment A Z
treatment B X
treatment B Y
treatment B Z
treatment C X
treatment C Y
treatment C Z
100
200
300
400
20
40
60
80

6. An experiment was performed to investigate the effect of plant density, variety and irrigation on the yield of carrots. There were four plant densities (20, 30, 40 and 50 plants per square metre) and three varieties (Autumn King, Chantenay, Nantes), with plots either irrigated or not irrigated. The experiment was arranged as a randomised complete block design in three blocks, each block containing 24 equal sized plots. Within each replicate block, the 24 treatment combinations were allocated at random to plots.

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

		Density (plants per square metre)			
Variety	Irrigate	20	30	40	50
Autumn King	No	66.5	75.3	92.3	95.4
Autumn King	Yes	125.9	164.0	193.9	211.6
Chantenay	No	54.5	67.5	87.3	90.6
Chantenay	Yes	124.3	151.7	189.0	199.5
Nantes	No	54.5	82.0	83.9	96.1
Nantes	Yes	120.6	151.2	180.9	186.0
Totals		546.3	691.7	827.3	879.2

	Irrigate	No	Yes	Totals
Variety	Autumn King	329.5	695.4	1024.9
	Chantenay	299.9	664.5	964.4
	Nantes	316.5	638.7	955.2
Totals		945.9	1998.6	2944.5

The three block totals (for 24 plots each) are 1021.7, 966.8 and 956.0, and the sum of squares for the 72 observations is 140 723.63. You may also use the facts that the sum of squares for the 24 treatment totals ($66.5^2 + 75.3^2 + \ldots + 186.0^2$) is 420 867.43, that the sum of squares for the twelve totals for Variety/Density combinations is 745 701.17, and that the sum of squares for the eight totals for Irrigation/Density combinations is 1 260 235.05.

(i) Construct an analysis of variance to assess the effects of plant density, variety and irrigation (and the interactions between these factors) on the yield of carrots. (7)

(ii) Extract the sums of squares for the linear and quadratic single-degree-of-freedom contrasts for the density factor [the coefficients of the linear contrast for four equally-spaced levels of a factor are (–3, –1, 1, 3) and those for the quadratic contrast are (–1, 1, 1, –1)]. (3)

(iii) Draw a diagram showing the treatment means for all 24 combinations of variety, irrigation and density, illustrating how the effects of density treatments vary with variety and irrigation. (4)

(iv) Using the diagram and the analysis of variance, explain the results found by this experiment, including the quantitative response to changes in plant densities, and any important interactions between the three treatment factors (variety, irrigation and plant density). (6)

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

(5)

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

(ii) Identify an appropriate confounding scheme to allow estimation of all main effects and two-factor interactions in a confounded factorial design for this experiment, clearly showing the contents of the principal block. From this write down also the contents of the other 7 blocks.

(6)

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

(4)

(iv) Before starting the experiment, the experimenter returns to you for more advice, having identified two possible modifications of the experimental set up. Briefly describe how each possible modification (separately) would influence your approach to designing the experiment and the structure of the analysis of variance table.

(a) Practical constraints associated with the application of the treatments mean that factor F needs to be kept constant during each day (changing the levels of this factor requires the experimental equipment to be taken apart and re-built, which can only take place overnight).

(2)

(b) A miscalculation of the costs associated with each experimental unit means that the experimenter can only afford to include 32 experimental units, still with 8 units processed during each day, the experiment now being completed in 4 days.

(3)

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

(2)

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

(3)

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

(3)

A project concerned with studying the effect of altitude changes on systolic blood pressure collected data on 39 male individuals born at high altitude in a primitive environment in the Peruvian Andes, who had then moved into the mainstream of Peruvian society living at a much lower altitude. The study was concerned with assessing the hypothesis that migration from high altitude to lower altitude would result in increased blood pressure at first, with a subsequent decrease to normal levels after a period living at lower altitudes. The systolic blood pressure (*sbp*) was measured for each individual, together with eight possible explanatory variables – years since migration (*years*), *age* (in years), *weight* (in kg), *height* (in mm), chin skin fold (in mm – *chin*) forearm skin fold (in mm – *forearm*), calf skin fold (in mm – *calf*) and pulse rate (in beats per minute – *pulse*).

The output **on the next page** shows the results of using a backward elimination stepwise procedure with the variance ratios for the inclusion or exclusion of variables from the model both set to have the value 1.5, and the model summary and parameter estimates for the selected model. This is also the best-fitting model based on consideration of the adjusted R-squared statistic in an all possible subsets selection process.

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

(4)

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

(vi) Interpret the analysis output for the backward elimination stepwise process, including an interpretation of the estimated parameters. (4)

Output for Question 8 is on the next page

Backward 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: 105.5 Dropping pulse       105.9 Dropping calf       107.3 Dropping forearm
        108.5 Dropping age         108.9 No change          113.4 Dropping chin
        113.9 Dropping height      127.9 Dropping years     176.4 Dropping weight
Chosen action: Dropping pulse.

Step 2: 102.6 Dropping calf        104.0 Dropping forearm    105.3 Dropping age
        105.5 No change            108.9 Adding pulse        110.2 Dropping chin
        110.4 Dropping height      124.1 Dropping years      171.4 Dropping weight
Chosen action: Dropping calf.

Step 3: 100.8 Dropping forearm     102.5 Dropping age        102.6 No change
        105.5 Adding calf          105.9 Adding pulse        107.0 Dropping chin
        107.3 Dropping height      122.1 Dropping years      166.4 Dropping weight
Chosen action: Dropping forearm.

Step 4: 100.0 Dropping age         100.8 No change           102.6 Adding forearm
        103.9 Adding pulse         104.0 Adding calf         104.1 Dropping height
        108.6 Dropping chin        118.9 Dropping years      172.4 Dropping weight
Chosen action: Dropping age.

Step 5: 100.0 No change            100.8 Adding age          102.4 Dropping height
        102.5 Adding forearm       103.1 Adding calf         103.1 Adding pulse
        106.6 Dropping chin        130.7 Dropping years      168.2 Dropping weight
Chosen action: No change.
```

Source	df	SS	MS	MS ratio
Regression	4	3130	782.5	7.82
Residual	34	3402	100.0	
Total	38	6531	171.9	

Percentage variance accounted for 41.8
Standard error of observations is estimated to be 10.0.

Parameter	estimate	s.e.	t value	p-value
Constant	107.2	51.0	2.10	0.043
chin	-1.410	0.778	-1.81	0.079
height	-0.0511	0.0378	-1.35	0.186
weight	1.879	0.377	4.98	<.001
years	-0.646	0.189	-3.43	0.002

BLANK PAGE