THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 4

SPECIMEN PAPER A

SOLUTIONS ARE CONTAINED IN A SEPARATE FILE

The time for the examination is 3 hours. The paper contains eight questions, of which candidates are to attempt **five**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

Note. In accordance with the convention used in all the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

1. (i) (a) State the *Gauss-Markov* theorem.

(2)

(b) State the conditions under which you might consider using weighted least squares rather than ordinary least squares in simple linear regression.

(4)

(ii) Data were collected about the performance of children in 14 schools. The children took an aptitude test before going to the schools, and then they sat a national test in mathematics five years later. The researchers expected there to be a relationship between a child's scores in these two tests. The mean scores were presented for each of the schools, and are given in the table below.

School	*Number of pupils*	*Mean score on aptitude test*	*Mean score on mathematics test*
1	12	70.6	44.6
2	35	61.5	51.0
3	24	67.3	59.2
4	10	79.3	85.7
5	32	65.2	53.6
6	25	64.3	56.1
7	21	67.5	42.8
8	20	60.2	52.2
9	15	66.0	39.1
10	7	80.5	74.7
11	27	52.4	41.6
12	37	70.4	40.4
13	19	64.2	53.7
14	27	68.4	39.6

(a) Draw a scatter plot, describe the data, and comment on any apparent relationship between the schools' mean scores from the two tests.

(6)

(b) The two sets of output **on the next page** give the results of ordinary linear regression and weighted linear regression, where the weights are the numbers of pupils from each school. Compare and contrast the two analyses and state which one you think is more appropriate, justifying your answer.

(8)

Output for question 1 is on the next page

Turn over

Output for question 1

Regression Analysis

The regression equation is
maths = -26.1 + 1.17 aptitude

Predictor	Coef	Stdev	t-ratio	p
Constant	-26.14	29.14	-0.90	0.387
aptitude	1.1732	0.4327	2.71	0.019

R-sq = 38.0% R-sq(adj) = 32.8%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	1	921.4	921.4	7.35	0.019
Error	12	1504.2	125.3		
Total	13	2425.6			

Weighted analysis using number of children as weight

The regression equation is
maths = 8.1 + 0.636 aptitude

Predictor	Coef	Stdev	t-ratio	p
Constant	8.09	30.32	0.27	0.794
aptitude	0.6363	0.4611	1.38	0.193

R-sq = 13.7% R-sq(adj) = 6.5%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	1	4413	4413	1.90	0.193
Error	12	27809	2317		
Total	13	32222			

Turn over

2. Large software packages comprise a large number of modules, or subroutines. The modules undergo careful testing to remove any "bugs". Various specification parameters are recorded for a random selection of modules to assess how the numbers of faults vary with the specification parameters. The data in the table below record the numbers of detected faults, y, together with two global specification parameters, $SPEC1$ and $SPEC2$, for 12 software modules written by the same programmer.

y	$SPEC1$	$SPEC2$
3	14.154	0.1132
10	31.817	4.5437
4	2.203	5.1989
24	22.646	15.0614
5	8.585	2.6844
4	2.160	11.2151
43	53.517	22.5853
3	6.234	0.7164
2	2.858	0.8493
26	34.124	16.0000
3	2.484	5.6245
2	6.619	0.1385

A generalised linear model is fitted to the data using Poisson errors and a log link function.

(i) Explain why such a distribution and link function might be appropriate for these data.

(1)

(ii) A model is fitted with

$$\eta = \beta_0 + \beta_1 SPEC1 + \beta_2 SPEC2$$

The scaled deviance is 11.96. Comment on the apparent fit of the model to the data.

(2)

(iii) Figures 2.1 to 2.3 in the **output on the following pages** show plots of the Pearson residual against the predicted value and each of the predictor values. Give detailed comments on the form of these plots.

(5)

(iv) Figures 2.4 and 2.5 in the **output on the following pages** show a histogram and Normal plot of the Pearson residuals. Comment on these plots.

(4)

Question 2 is continued on the next page

(v) A second model was fitted with

$$\eta = \beta_0 + \beta_1 \log\left(SPEC1\right) + \beta_2\, SPEC2 \ .$$

Comment on whether you think this was a sensible idea.

(1)

(vi) The scaled deviance from this second model is 4.3478. Comment on the apparent fit of this model.

(2)

(vii) Figures 2.6 to 2.10 in the **output on the following pages** show residual plots and a histogram and Normal plot of the Pearson residuals from this second model. Comment on the plots.

(2)

(viii) What further analyses would you carry out? Justify your answer.

(3)

Output for question 2 is on the next four pages

Output for question 2. This output is printed on this page and the next three pages

Residual plots from model with spec1 and spec2

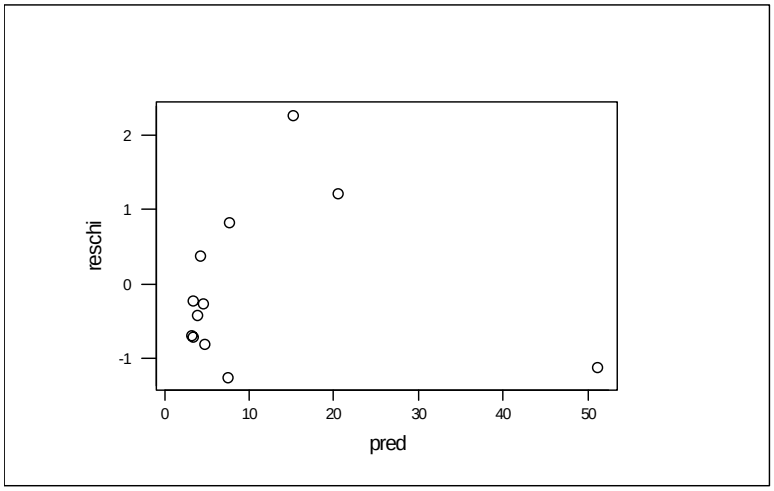

Figure 2.1. Residuals against predicted values

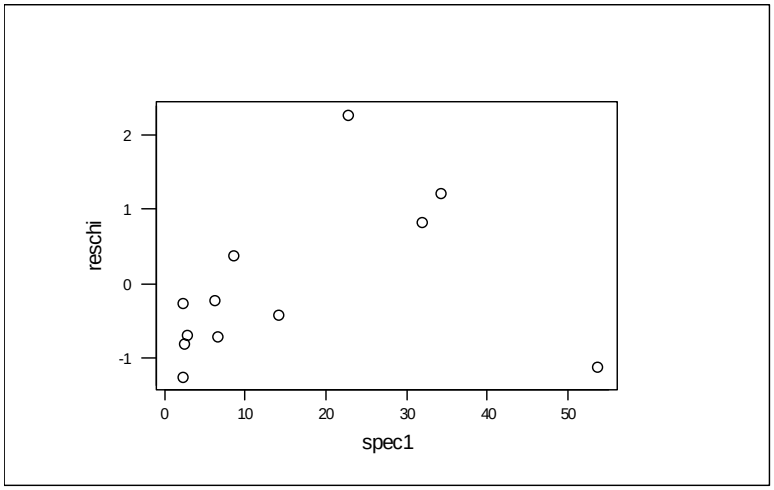

Figure 2.2. Residuals against SPEC1

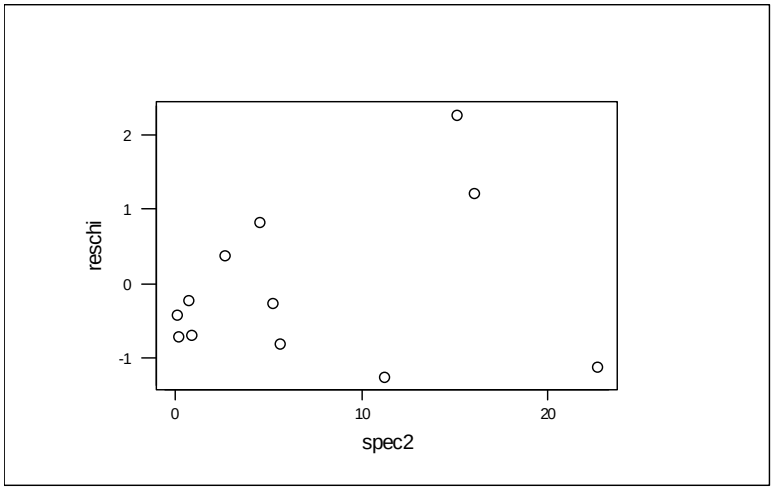

Figure 2.3. Residuals against SPEC2

Turn over

Output for question 2 (contd)

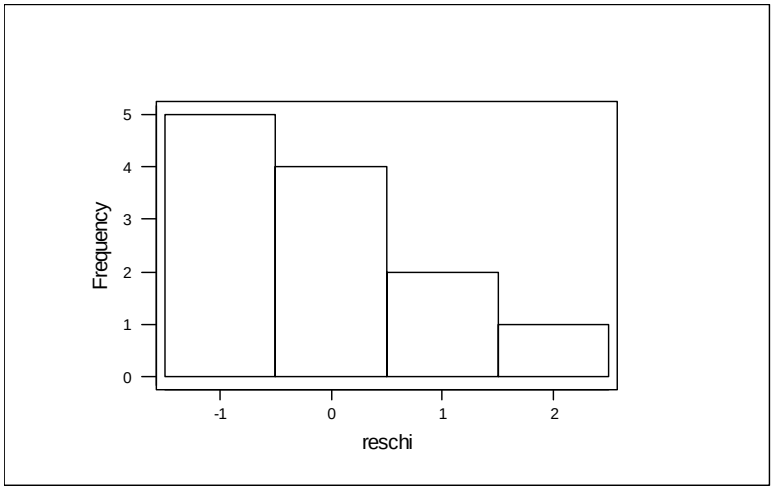

Figure 2.4. Histogram of residuals

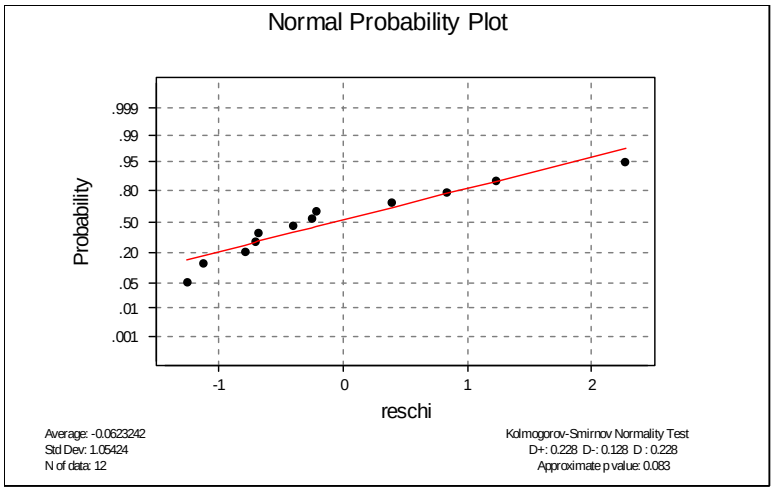

Figure 2.5. Normal plot of residuals

Turn over

Output for question 2 (contd)

Residual plots from model with log(spec1) and spec2

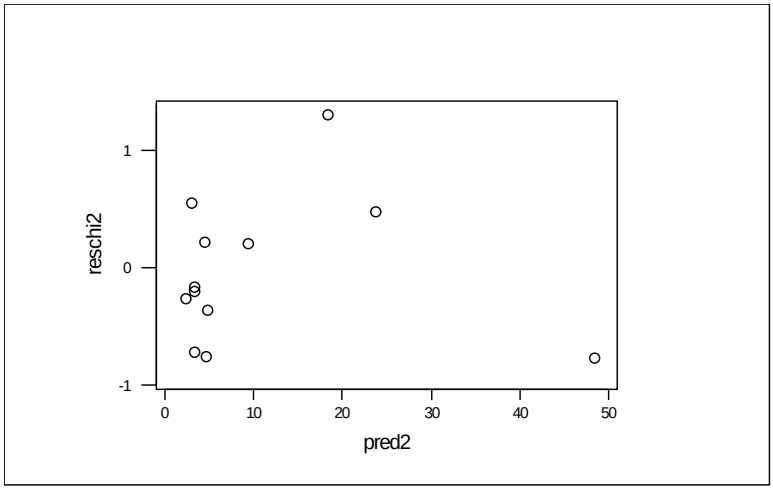

Figure 2.6. Residuals against predicted values

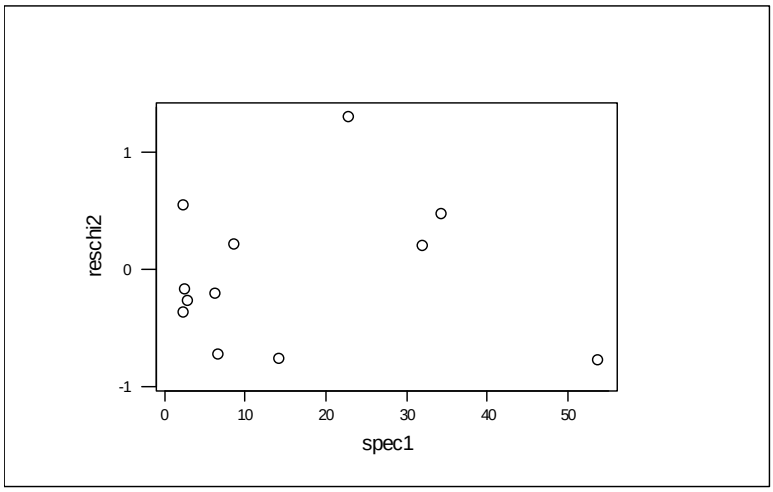

Figure 2.7. Residuals against SPEC1

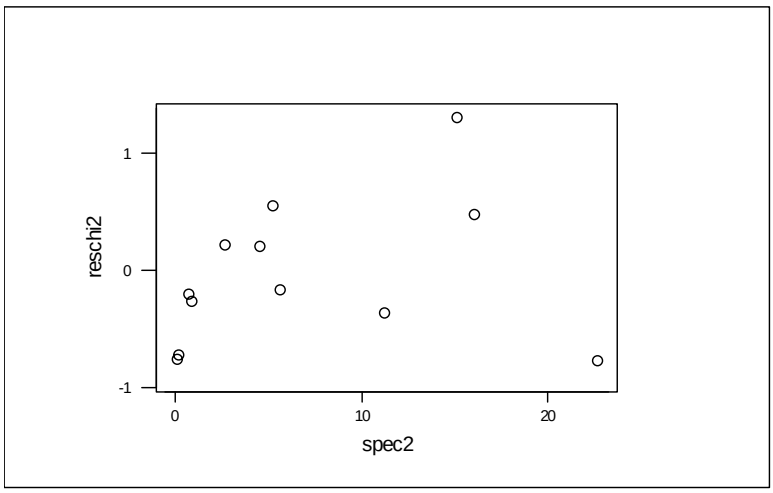

Figure 2.8. Residuals against SPEC2

Turn over

Output for question 2 (contd)

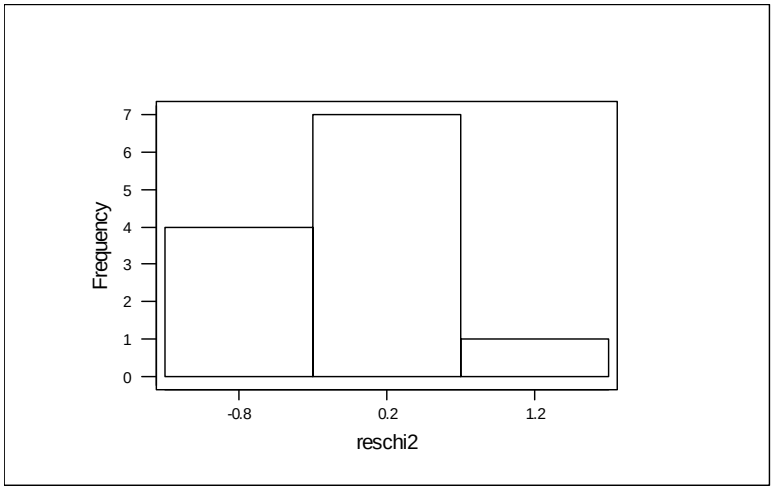

Figure 2.9. Histogram of residuals

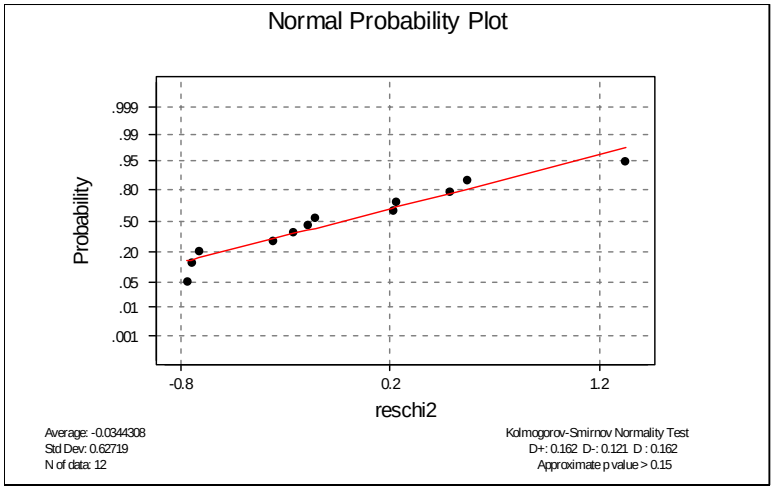

Figure 2.10. Normal plot of residuals

Turn over

3. A researcher is fitting a regression model predicting an observed variable y from another variable x, the values of which the researcher can control. The available data are given in the table, presented here in ascending order of values of x.

y	2.2	1.9	2.7	1.6	2.3	1.9	1.8	3.6	1.6	2.8	2.8
x	1.2	1.2	2.1	2.1	2.8	3.4	3.4	3.8	3.8	4.1	4.1

y	2.1	3.3	1.8	1.9	2.9	2.2	3.4	3.5	3.3	3.1	3.0
x	4.1	4.6	4.6	5.1	5.2	5.2	5.4	5.8	6.1	6.1	6.2

(i) Draw a scatter plot of the data. Using this, discuss the nature of any association between the variables and the nature of the variability in the data.

(5)

(ii) Explain how the information about the table helps to justify the researcher's decision to use a linear regression model to analyse the data, splitting the residual error term into two parts, for "lack of fit" and "pure error".

State the linear model underlying this analysis, and the properties of the terms in it.

(5)

(iii) (a) Show that the pure error SS from repeats at $x = 1.2$ is 0.045 and state the associated degrees of freedom.

(b) Calculate the pure error SS from repeats at $x = 4.1$ and state its associated degrees of freedom.

(c) Without further working state how the total pure error SS is computed.

(4)

(iv) Using the results of a linear regression, presented below, together with the fact that the pure error SS is 4.3717, carry out a lack of fit test, and state your conclusions.

(4)

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2.6723	2.6723	8.03	0.010
Residual Error	20	6.6554	0.3328		
Total	21	9.3277			

(v) Describe how lack of fit can be detected when there are no repeated observations of y at any x value.

(2)

4. (i) Briefly describe the advantages and disadvantages of the backward elimination method of model selection in multiple linear regression.

(4)

(ii) A dataset of 13 observations contains four predictor variables (X1, X2, X3, X4) and one response variable (Y), and the following statistics are available.

Variables in linear model	Model Sum of Squares
X1, X2, X3, X4	2667.90
X2, X3, X4	2641.95
X1, X3, X4	2664.93
X1, X2, X4	2667.79
X1, X2, X3	2667.65
X3, X4	2540.00
X2, X4	1846.88
X2, X3	2300.30
X1, X4	2641.00
X1, X3	1488.70
X1, X2	2657.90
X4	1831.90
X3	776.40
X2	1809.40
X1	1450.10
Total Sum of Squares	2715.76

Apply a backward elimination method to select the set of predictor variables that you consider "best" model the data.

(8)

(iii) Explain how your method of model selection might be different if you knew what the predictor variables were.

(4)

(iv) A journal gives the following advice to authors.

"Automated stepwise techniques often produce wildly unreliable results. This includes not only forward and backward automated selection but also 'best subset' approaches. Manuscripts that employ these techniques will not be considered unless the model is supported by a validation procedure."

(a) Do you agree with the first sentence? Justify your answer.

(b) Suggest possible validation procedures that could be used for the model chosen in part (ii).

(4)

5. An experiment was carried out to study brake wear (y) on cars of similar size made by four different manufacturers F, V, R and P. Four different test roads (1, 2, 3, 4) were used and there were four drivers (A, B, C, D). The scheme for the experiment, and the results y, are shown in the table.

		Driver and response (y)				
		Road 1	*Road 2*	*Road 3*	*Road 4*	*Total*
Car manufacturer	*F*	*B*: 44	*A*: 46	*D*: 39	*C*: 52	181
	V	*C*: 51	*B*: 37	*A*: 43	*D*: 40	171
	R	*A*: 42	*D*: 39	*C*: 46	*B*: 34	161
	P	*D*: 45	*C*: 52	*B*: 36	*A*: 42	175
	Total	182	174	164	168	

$\Sigma y = 688, \qquad \Sigma y^2 = 30\,042.$

(i) Analyse the data to extract effects due to manufacturers, roads and drivers, and report briefly on the results.

(6)

(ii) Discuss briefly the advantages and disadvantages of a Latin square design.

(4)

(iii) Explain the principles that should be followed when selecting a particular Latin square of the appropriate size for use in an experiment.

(4)

You are now told that drivers A and C always drove in the morning, while B and D always drove in the afternoon. Also, A and B drove at the weekend, while C and D drove on weekdays.

(iv) Investigate the data further, using linear contrasts to examine differences between times of day, differences between times in the week, and any other relevant comparisons. Discuss your results, and explain whether it is reasonable to ascribe these differences to time.

(6)

6. The germination rate of a particular species of plant is studied under four different growing conditions. The four conditions are the factorial arrangements of two storage methods, stratified and unstratified, and two storage temperatures, called "spring" and "summer".

Forty batches of seeds were prepared, each containing 20 seeds, and 10 batches were allocated at random to each of the four treatments. First the seeds were stored for two weeks according to their pre-assigned storage method. Then the seeds were placed on dishes with 5 ml of water, and stored for two weeks according to their pre-assigned storage temperature.

The table below shows the numbers of germinated seeds (out of 20) for each batch of seeds, and also (in brackets) the residuals obtained from a one-way analysis of variance.

Spring/ Stratified	12	13	2	7	19	0	0	3	17	11
	(3.6)	(4.6)	(–6.4)	(–1.4)	(10.6)	(–8.4)	(–8.4)	(–5.4)	(8.6)	(2.6)
Spring/ Unstratified	6	2	0	2	4	1	0	10	0	0
	(3.5)	(–0.5)	(–2.5)	(–0.5)	(1.5)	(–1.5)	(–2.5)	(7.5)	(–2.5)	(–2.5)
Summer/ Stratified	6	4	5	7	6	5	7	5	2	3
	(1.0)	(–1.0)	(0.0)	(2.0)	(1.0)	(0.0)	(2.0)	(0.0)	(–3.0)	(–2.0)
Summer/ Unstratified	0	6	2	5	1	5	2	3	6	6
	(–3.6)	(2.4)	(–1.6)	(1.4)	(–2.6)	(1.4)	(–1.6)	(–0.6)	(2.4)	(2.4)

(i) A one-way analysis of variance was conducted, to compare the four treatments. Explain in detail how the residuals shown above were calculated. Explain how you would analyse the residuals to examine the usual assumptions concerning Normality and constant variance.

(14)

(ii) Write down all the assumptions required for this analysis of variance. Discuss whether you would expect these data to satisfy all the assumptions, and explain any concerns you have about the validity of the analysis. State the steps which could be taken to overcome any concerns you have.

(6)

7. An experiment was carried out to investigate both the effect of rate of seeding and the effect of spatial arrangement on the yield of turnips. Five seeding rates (0.5, 2, 8, 20, 32 lb/acre) and four row widths (4, 8, 16, 32 inches) were tested. The experiment was laid out in three blocks, each with 20 equal-sized plots. Within each block, the 20 treatment combinations were allocated at random to the plots.

The table below summarises the total yields for the three plots for each treatment combination (in coded units), obtained during a fixed period in the growing season.

		Row width (inches)				
		4	8	16	32	*Total*
Seed rate (lb/acre)	0.5	1.87	3.07	3.13	2.85	10.92
	2	5.40	6.28	6.08	5.73	23.49
	8	7.67	7.94	6.94	6.58	29.13
	20	8.16	8.46	8.43	6.80	31.85
	32	8.38	8.40	7.81	7.05	31.64
	Total	31.48	34.15	32.39	29.01	127.03

Block totals (of 20 plots each) are: I, 36.09; II, 43.27; III, 47.67.

The sum of the squares of all 60 observations is 301.4107.

You may also use the fact that $1.87^2 + 5.40^2 + \ldots + 7.05^2 = 889.5165$.

(i) Carry out an analysis of variance to examine the effects of seed rate, row width, and their interaction, on the yield of turnips.

(6)

(ii) Partition the sum of squares for the row width main effect into single-degree-of-freedom components. Examine and comment on these.

[Note. The row width levels used were equally spaced on the logarithmic scale. The coefficients of linear, quadratic and cubic components for four equally-spaced levels of a factor are, respectively, (–3, –1, 1, 3), (1, –1, –1, 1) and (–1, 3, –3, 1).]

(5)

(iii) Draw a diagram showing all 20 totals of seed rate and row width combinations.

(4)

(iv) Using the diagram and the analysis of variance, explain the results found by this experiment, including mention of any interaction between seed rate and row width.

(5)

8. (a) The yields of two aubergine varieties, V_1 and V_2, were examined in two soil types, S_1 and S_2, and at two levels of moisture, M (low, M_1, and high, M_2). Four complete replicates of a 2^3 factorial design were run in 4 greenhouses. The yields y (coded in suitable units) and treatment combinations are given below.

In the coding of the treatment combinations,
presence of v indicates V_2 was used, otherwise V_1 was used;
presence of s indicates S_2 was used, otherwise S_1 was used;
presence of m indicates M_2 was used, otherwise M_1 was used.

	Greenhouse				
	I	**II**	**III**	**IV**	Total
(1)	7	19	13	11	50
v	30	33	28	31	122
s	24	30	19	25	98
vs	39	36	35	43	153
m	21	30	24	21	96
vm	31	36	31	33	131
sm	27	31	26	29	113
vsm	39	41	36	66	182
Total	218	256	212	259	

Effect estimates	
V	14.44
S	9.19
M	6.19
VS	1.06
VM	–1.44
SM	–0.69
VSM	3.19

$\Sigma y = 945 \qquad \Sigma y^2 = 31513$

(i) Explain briefly why the above is called a factorial layout. What are the advantages of this type of design?

(3)

(ii) Show how the value 3.19 for the *VSM* interaction estimate was obtained.

(2)

(iii) Construct the analysis of variance. Carry out any further significance tests that you consider appropriate, and report on the results.

(8)

(b) Suppose now that only four aubergine plants can be grown in each greenhouse, and that the greenhouse used may influence the yield. However, 8 greenhouses of this size are available.

(i) Explain briefly the importance of *confounding* in 2^k factorial experiments.

(3)

(ii) Write down an appropriate design for an experiment using the same treatment combinations as in part (a), in which each of these 8 greenhouses is treated as a block.

(4)