THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 4

SPECIMEN PAPER B

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) Explain what is meant by a *transformation to stabilise variance*, and give an example of where this might be useful in linear regression.

(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 expansion to derive approximate expressions for $E(X)$ and $\mathrm{Var}(X)$ in terms of the expected value and variance of Y. Hence, or otherwise, show that if σ is a function $f(\mu)$ of the mean μ, an appropriate variance-stabilising transformation for Y might be $h(Y)$ defined through the relationship

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

(3)

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

The standard deviation is proportional to the mean.
The standard deviation is proportional to the square of the mean.
The variance is proportional to the mean.

(3)

(iv) An experiment was performed to examine the relationship between calcium and strength of fingernails. A sample of 32 students was selected and they were given calcium supplements in amounts of either 10mg, 20mg, 30mg or 40mg. Eight students were allocated at random to each amount. After a set period, fingernail strength was measured. The results are given in the table.

10mg	*20mg*	*30mg*	*40mg*
13	114	48	79
42	75	72	314
28	27	104	75
12	136	133	108
67	100	57	90
62	83	112	175
28	56	87	108
29	37	64	74

Calculate suitable summary statistics and hence sketch a graph to help to examine possible transformations to stabilise variance for these data. Describe the practical problems of applying the theory from part (iii) when analysing the results of this experiment.

(5)

(v) Discuss how you would proceed to fit a suitable model. How would you decide between models that had the predictor variable (calcium supplement amount) as a covariate or a factor?

(6)

Turn over

2. Suppose that independent random variables $Y_1, Y_2, \ldots, Y_n$ follow binomial distributions, that is

$$Y_i \sim \mathrm{B}(m_i, \pi_i),\ i = 1, 2, \ldots, n, \quad \text{where } P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}.$$

(i) (a) Show that the natural canonical link function for this distribution in the context of generalised linear models is given by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

where $\eta_i = \mathbf{x_i}^T \boldsymbol{\beta}$ is the linear predictor.

(2)

(b) Define the terms *odds* and *log-odds* in the context of such a model. Explain how these values may be estimated after such a model has been fitted. Explain also how approximate 95% confidence intervals may be calculated for these quantities, given that the necessary standard errors for the linear predictor values are available.

(3)

(ii) The following data show the perinatal mortality rates for children born to a group of women classified by their age and the length of the gestation period.

Gestation period (days)	*Mother's age (years)*	*Mortality/Total Births*
197 – 260	<30	59/414
	≥30	45/203
>260	<30	30/4501
	≥30	15/1633

The data were analysed by fitting a binomial model with logit link. The variables used were:

GEST coded as 0 for 197 – 260, and 1 for >260

AGE coded as 0 for <30, and 1 for ≥30.

The results are summarised below.

Question 2 is continued on the next page

Variables in model	*Scaled deviance*	*Parameter estimate*	*(standard error)*
constant	346.25	–3.7912	(0.0828)
constant AGE	334.08	–3.9931 0.6054	(0.1070) (0.1693)
constant GEST	6.88	–1.5959 –3.3117	(0.1075) (0.1843)
constant AGE GEST	0.3140	–1.7659 0.4677 –3.2886	(0.1296) (0.1798) (0.1846)

(a) Briefly describe two situations, one where the method of sampling would be appropriate to using the above model for the analysis of these data, and one where it would not. Justify your answers.

(3)

For the rest of this question, you should assume that it is valid to use the above model.

(b) Using forward selection, discuss with reasons which combination of variables in the model best describes the data.

(4)

(c) Using the output for your chosen combination, estimate the odds of mortality for the group "mother's age less than 30 years and gestation period 197 – 260 days"; estimate also the probability of mortality for this group. Calculate approximate 95% confidence intervals for both these quantities.

(5)

(d) Calculate the odds ratio for the mortality in the group in (c) compared to the "less than 30 years and at least 260 days" group, and give an approximate 95% confidence interval for this quantity.

(3)

3. (i) Give a brief description of the *forward selection procedure* for selecting a multiple linear regression model. Your description should include the following.

Reasons for using model selection procedures.
The relative merits of the method.
The steps required to carry out the procedure.

(6)

(ii) A metal alloy was being developed to have a prescribed percentage elongation under stress. Varying amounts of additives *X1*, *X2* and *X3* were used to produce 24 experimental pieces of alloy, and their elongations were measured. Multiple linear regression models were fitted as shown in the table below.

Terms in model (in addition to intercept)	*Error sum of squares*
–	170.85
X1	165.97
X2	117.17
X3	122.43
X1, X2	116.30
X1, X3	121.75
X2, X3	90.007
X1, X2, X3	88.453

(a) Using the forward selection procedure, find the model which best explains the data. At each step, state clearly the hypothesis being tested, the value of the test statistic and the outcome of the test, specifying an appropriate significance level.

(7)

(b) Mallows' C_p statistic for a model containing s parameters is defined by

$$C_p(s) = \frac{SS_E}{\hat{\sigma}^2} - (n - 2s)$$

where SS_E is the error sum of squares for the model and $\hat{\sigma}^2$ is the variance estimated from the full model (which is the final model in the table above). Calculate $C_p(s)$ for each of the models in the table and comment, referring to desirable values of $C_p(s)$. Does the full model appear to provide the best fit to the data? Justify your answer.

(7)

Turn over

4. (a) A property investor is trying to model current property prices in terms of the age and type of property. He has data on 55 properties for sale in his area. These data consist of price in thousands of UK pounds, age of property in years, and type of property. The type is represented by the following coding:

detached house = 1
semi-detached house = 2
terraced house = 3.

(i) Contrast the terms *factor* and *continuous variable* as used in linear modelling.

(2)

(ii) The property investor has done some linear modelling, using multiple regression, with both age and type as predictor variables. Part of the output is shown below. Say, with reasons, whether "type" has been coded as a factor or a continuous variable. Comment on whether you think this is a sensible choice.

(4)

```
Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    2      5717.5      2858.7      6.56    0.003
Error        52     22662.3       435.8
Total        54     28379.8
```

(iii) After receiving statistical advice, he runs a further multiple regression analysis using two different packages, and appears to get different output.

Data for the first three observations are **shown on the next page**, together with the coefficients and the first three rows of the design matrix for package A. Part of the output from package B is also shown.

Write down the first three rows of the design matrix used in package B. Show that the parameter estimates obtained by the two packages are consistent with each other. State any other ways in which the two sets of output would differ.

(9)

Question 4 is continued on the next page

Sample Data (first three observations)

Observation	Price	Age	Type
1	54.0	58	3
2	54.3	19	2
3	55.2	10	1

Output from package A

```
Parameter estimates:
78.8603   11.2249   -0.5764   -0.4180

First 3 rows of Design Matrix
   1   -1   -1   58
   1    0    1   19
   1    1    0   10
```

Output from package B

```
Intercept     68.212
Age           -0.418
[type=1]      21.873
[type=2]      10.072
[type=3]       0
```

(b) A health psychologist is studying fear of falling in old people. He is examining the relationships between three psychometric scales using a random sample of 64 people. Scale A measures fear of falling, where a high score corresponds to high fear. Scale B measures confidence doing tasks where there is a risk of falling; here a high score corresponds to high confidence. Scale C is a measure of general anxiety, where a high score corresponds to high anxiety.

Pearson correlations from his sample are as follows.

	Scale A	Scale B
Scale B	-0.771	
Scale C	0.637	-0.328

Coefficients from multiple regression, modelling Scale B as the dependent variable, are as follows.

Dependent variable: Scale B

	Coefficient	Standard error
Constant:	126.28	2.976
Scale A:	-1.486	0.165
Scale C:	1.439	0.566

Given the negative correlation between Scales B and C, the psychologist expected the coefficient of Scale C to be negative. Provide a possible explanation for the positive coefficient.

(5)

5. A manager in a factory knows little about statistics. On the basis of statistical advice, he carried out an experiment to decide which of three machines was best to buy. Six employees were chosen at random. Each employee operated each machine twice. Each piece of work was rated on its quality, where a high score is better than a low score. The manager wanted to choose a machine on which employees would produce high quality work.

The 36 runs in the experiment were carried out in random order.

The data are given below, together with some related statistics.

Machine	*Employee*	*Rating*	*Machine*	*Employee*	*Rating*	*Machine*	*Employee*	*Rating*
1	1	52.4	2	1	64.5	3	1	67.5
1	1	50.4	2	1	63.8	3	1	67.2
1	2	51.9	2	2	59.4	3	2	61.4
1	2	52.6	2	2	59.2	3	2	61.9
1	3	60.2	2	3	55.0	3	3	70.5
1	3	61.5	2	3	65.7	3	3	70.6
1	4	51.3	2	4	62.4	3	4	64.9
1	4	52.4	2	4	62.3	3	4	66.5
1	5	64.5	2	5	64.9	3	5	72.4
1	5	63.5	2	5	65.1	3	5	71.3
1	6	46.6	2	6	43.8	3	6	62.5
1	6	49.5	2	6	44.6	3	6	60.4

MEANS

Machine	N	Rating
1	12	54.733
2	12	59.225
3	12	66.425

Employee	N	Rating
1	6	60.967
2	6	57.733
3	6	63.917
4	6	59.567
5	6	66.950
6	6	51.233

TOTALS

	Employee						
Machine	1	2	3	4	5	6	All
1	102.8	104.5	121.7	103.7	128.0	96.1	656.8
2	128.3	118.6	120.7	124.7	130.0	88.4	710.7
3	134.7	123.3	141.1	131.4	143.7	122.9	797.1
All	365.8	346.4	383.5	359.8	401.7	307.4	2164.6

(Note: $102.8^2 + 104.5^2 + \ldots + 122.9^2 = 264308.12$.

You may also use the fact that the corrected total sum of squares for all 36 observations is 2071.99.)

(i) Write down an appropriate model for this design, explaining all terms and stating the necessary assumptions. Indicate which (if any) of the terms are "fixed effects" and which (if any) are "random effects". (6)

(ii) State the expected mean squares of the effects in your model. (3)

(iii) Complete an analysis of variance for these data, and write a report on the results. (You may use the tables of means and totals above as required.) (11)

6. Explain clearly what is meant by a *balanced incomplete block* design. When is this design useful?

(4)

Seven different confectionery products, A to G, made from the same ingredients but using slightly different recipes, were examined by a panel of experts. There were 7 panel sessions; at each session, 3 recipes were tested, the order of testing being random. The panel assessed the recipes blind, and gave a total score to each recipe based on a variety of characteristics. A higher score indicated a better perceived quality.

The scheme for the experiment and the results (y) were as shown in the table.

Block (panel session)	I	II	III	IV	V	VI	VII
Product tested and response y	A: 20	B: 25	C: 24	A: 16	B: 20	C: 19	A: 19
	B: 23	D: 21	E: 18	D: 14	E: 16	D: 17	F: 20
	C: 16	F: 20	F: 19	E: 15	G: 25	G: 22	G: 24
Session total	59	66	61	45	61	58	63

$\sum y = 413$, $\sum y^2 = 8341$.

The estimated treatment effects adjusted for blocks ($\hat{\tau}_i$) were:

(Adjusted) estimated treatment effects						
A	B	C	D	E	F	G
–0.2857	2.5714	–0.1429	–1.8571	–2.8571	–1.8571	4.4290

(i) Verify that this is a balanced incomplete block design with panel sessions forming the blocks, and find the values of its structural parameters N, b, k, v, r and λ. Define all the notation that you use.

(5)

(ii) Use the estimated treatment effects (adjusted for blocks) given above to construct the analysis of variance.

[You may use the fact that the treatment sum of squares, adjusted for blocks, is $\frac{v\lambda}{k}\sum \hat{\tau}_i^2$.]

(6)

(iii) Carry out any tests of significance which you consider necessary to investigate differences between treatments, and state your conclusions.

[Note. $\mathrm{Var}(\hat{\tau}_i - \hat{\tau}_j)$ can be estimated by $2k\hat{\sigma}^2/(v\lambda)$ for any $i \neq j$, where $\hat{\sigma}^2$ is suitably defined.]

(5)

Turn over

7. An experiment was conducted to compare several treatments, each of which was replicated *r* times during the experiment. Explain what is meant by a *contrast* in the comparison of treatment means and derive its standard error. Define any notation you use in your answers. State the conditions under which two contrasts are *orthogonal* and explain the relevance of orthogonality.

(6)

An experiment was conducted on the effect of inoculating *Phaseolus vulgaris* (bean) seeds with nitrifying *Rhizobium* bacteria. The aim of the experiment was to investigate the effect of a liquid fertiliser when used with different strains of Rhizobium. Eight treatments were used, labelled A – H, and defined below. The fertiliser was used at two levels. Treatments G and H were "controls", not inoculated with Rhizobium.

Treatment	*Strain of Rhizobium*	*Cultural history*	*Fertiliser level*
A	R 3644	Newly-cultured	Low
B	R 3644	Newly-cultured	High
C	R 3644	Repeatedly subcultured	Low
D	R 3644	Repeatedly subcultured	High
E	CC 511	Peat-based	Low
F	CC 511	Peat-based	High
G	Non-inoculated		Low
H	Non-inoculated		High

The experiment consisted of 5 replicates in a completely randomised design. The response was total root nodule weight after a fixed period of time.

The mean responses (in coded units) for the 5 replicate samples were:

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
88	198	66	235	265	233	40	41

(i) Write down a set of meaningful orthogonal contrasts which assess the following types of treatment differences:

(a) effect of fertiliser;

(b) difference between the two cultures of R 3644;

(c) effect of inoculation;

(d) difference between the two strains of Rhizobium;

(e) interactions of (b), (c) and (d) with fertiliser.

(6)

(ii) Calculate the value of each contrast. Given that the residual (error) mean square is 3265.8, test the statistical significance of each contrast, stating any assumptions required for the validity of the test. Summarise the results found by this experiment, including mention of any fertiliser/*Rhizobium* interactions.

(8)

8. An experiment on the growth of tomato plants is carried out in a glasshouse. The plants are grown in commercial "grow-bags" instead of in pots. Each grow-bag contains a standard growing medium and will hold four plants, all of which must receive the same treatment.

There is sufficient space in the glasshouse to place on benches two rows of grow-bags, side by side with 16 in each row. The experiment includes three factors *A*, *B*, *C* that are different nutrients, with each factor used at two levels (low and high). The glasshouse length runs north to south and there are doors at each end.

(i) At the planning stage it was decided to include blocking in the design, with the northernmost eight bags forming block I and the remaining three blocks constructed in a similar way. With the aid of a sketch, explain why this decision would be better than complete randomisation. Show in your sketch how the treatments could be allocated in a typical block, using the coding in the following table of data.

(5)

(ii) The data below are the total yields (kg) of the four plants in each unit plot, obtained during a fixed period in the growing season. The treatment combinations are coded in the usual way; for example the letter *a* is included in the combination when factor *A* is present at its high level, and not otherwise.

		Treatment combination								*Block total*
		(1)	*a*	*b*	*ab*	*c*	*ac*	*bc*	*abc*	
Block	I	4.4	10.7	3.9	13.4	5.0	10.3	6.2	14.9	68.8
	II	7.6	10.8	5.0	19.0	6.4	13.4	4.0	15.6	81.8
	III	7.1	11.6	5.6	18.0	7.0	14.0	3.6	16.4	83.3
	IV	5.2	9.3	4.6	12.9	5.6	11.0	5.7	15.8	70.1
Treatment total		24.3	42.4	19.1	63.3	24.0	48.7	19.5	62.7	

The sum of squares of all 32 observations is 3563.28

(a) Copy and complete the analysis of variance table **shown on the next page**.

(6)

(b) Carry out such tests of significance as you consider necessary, and construct any diagrams that will help in interpreting the results.

(6)

(c) Write a brief report for the scientist who commissioned the experiment.

(3)

The analysis of variance table for part (ii)(a) is shown on the next page

Turn over

Analysis of variance table for question 8 part (ii)(a)

Analysis of Variance for total yield of 4 tomato plants per unit plot

Source of variation	DF	Sum of squares	
Blocks	3		*******
A	1	*******	
B	**	19.84500	
C	**	1.05125	
AB	**	62.16125	
AC	**	0.98000	
BC	**	1.20125	
ABC	**	*******	
Treatments	**		616.795
Residual	**		*******
Total	31		675.280