THE ROYAL STATISTICAL SOCIETY

2010 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

MODULE 4

MODELLING EXPERIMENTAL DATA

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many 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 these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

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

Graduate Diploma, Module 4, 2010. Question 1

(i) The systematic variation that is to be eliminated may arise from "days" and from "order". In the Latin square design, these are assigned to the columns and rows (or vice versa) of the square with the treatments (A, B, C, D, E) appearing within the square. Each of these letters appears exactly once in each row and in each column. Thus, if there are any systematic differences between the rows, they will affect each of the treatments equally; and likewise for the columns. The rows, columns and treatments are orthogonal in this sense. It has to be assumed that there is no interaction between rows, columns and treatments.

The usual linear model for analysing data from a Latin square experiment is

$$y_{ijk} = \mu + r_i + c_j + t_k + \varepsilon_{ijk}$$

for $i, j, k = 1, 2, ..., n$ [n being the size of the square, ie 5 in this example], noting that not all combinations occur, where

y_{ijk} = observation in row i and column j on treatment k

μ = overall grand population mean for the whole experiment

r_i = population mean effect for row i, measured as the difference from μ

c_j = population mean effect for column j, measured as the difference from μ

t_k = population mean effect for treatment k, measured as the difference from μ

ε_{ijk} are the residual (error) terms, assumed to be independent N(0, σ^2) random variables

and with $\Sigma r_i = \Sigma c_j = \Sigma t_k = 0$.

(ii) Standard Latin squares have the letters in alphabetical order in the first row and in the first column. One of these squares is chosen at random. The rows of this square are then permuted at random, as are the columns, to give a randomised design. The letters *A, B, C, D, E* are then allocated at random to the treatments. This gives a random choice from all possible 5×5 squares.

Solution continued on next page

(iii) Row totals are: 35.35, 37.23, 37.07, 37.58, 39.92.

Column totals are: 37.05, 37.79, 36.80, 37.70, 37.81.

Treatment totals are: A 35.59, B 44.15, C 31.44, D 31.88, E 44.09.

"Correction term" is $\frac{187.15^2}{25} = 1401.0049$.

Therefore total SS = 1436.4189 – 1401.0049 = 35.414.

SS for rows (orders)

$$= \frac{35.35^2}{5} + ... + \frac{39.92^2}{5} - 1401.0049 = 1403.1486 - 1401.0049 = 2.144.$$

SS for columns (days)

$$= \frac{37.05^2}{5} + ... + \frac{37.81^2}{5} - 1401.0049 = 1401.1825 - 1401.0049 = 0.178.$$

SS for treatments

$$= \frac{35.59^2}{5} + ... + \frac{44.09^2}{5} - 1401.0049 = 1432.9213 - 1401.0049 = 31.916.$$

Hence:

SOURCE	DF	SS	MS	*F* value
Rows (orders)	4	2.144	0.536	5.47 highly significant
Columns (days)	4	0.178	0.045	0.46 not significant
Treatments	4	31.916	7.979	81.42 very highly sig
Residual	12	1.176	0.098	$= \hat{\sigma}^2$
TOTAL	24	35.414		

[All *F* values are compared with $F_{4,12}$; upper 5% point is 3.26, upper 1% point is 5.41, upper 0.1% point is 9.63.]

The variance of the difference between any pair of treatment means is estimated by $\frac{2\hat{\sigma}^2}{5} = \frac{2 \times 0.098}{5} = 0.0392$. So the standard error of a difference is $\sqrt{0.0392} = 0.198$.

The means for treatments D and E are 6.38 and 8.82 respectively. So the difference (measured as E – D) is 2.44. The double-tailed 5% point of t_{12} is 2.179, so a 95% confidence interval for the true difference in means (E – D) is given by $2.44 \pm (2.179 \times 0.198) = 2.44 \pm 0.43$, i.e. it is (2.01, 2.87).

Solution continued on next page

As the confidence interval just found does not contain zero, we have evidence that the true mean effects for treatments D and E differ. Similar analyses may be carried out for the other pairs of treatment means. The overall analysis of variance has shown extremely strong evidence that there are real differences among these means. An informal inspection of them, or of the totals that are given in the question, might suggest that C and D are similar but give a lower result than A; and that B and E are similar, giving a higher result than A.

The question asks specifically for brief comment on the day and order terms. There is no evidence of any overall differences among the days. There is strong evidence that there are overall differences among the orders. This can be investigated in detail via *t* tests, but again an informal inspection of the totals (these are the row totals) might suffice, as it suggests quite strongly that the second, third and fourth orders are similar in their effects, giving a higher result than the first order and a lower result than the fifth order.

Graduate Diploma, Module 4, 2010. Question 2

(i) There are enough units for two complete replicates of the experiment, but they still need grouping into blocks of 8.

The first step for the design suggested below is therefore to conduct the full 2^5 experiment in four blocks of size 2^3. To do this, two contrasts must be selected for confounding, and their "product" will also be confounded. The best type of scheme for this situation is to select two 3-factor interactions which have only one letter in common, such as ABC and ADE; their product is the 4-factor interaction BCDE. Thus the confounding scheme is given as follows, with the generators underlined:

I ABC ADE BCDE.

In the usual notation for two-factor experiments, the principal block for this design is

(1) *bc* *de* *bcde* *acd* *abd* *ace* *abe* .

The other blocks are

"multiply" by *a*: *a* *abc* *ade* *abcde* *cd* *bd* *ce* *be*

"multiply" by *b*: *b* *c* *bde* *cde* *abcd* *ad* *abce* *ae*

"multiply" by *ab*: *ab* *ac* *abde* *acde* *bcd* *d* *bce* *e* .

This gives a layout for 32 plots of the experiment, and the same layout would be repeated for the other 32. [Note. A case could be made for using a different confounding scheme for the other 32 so as to reduce the number of confounded interactions in the whole experiment. Candidates were not expected to discuss this in the examination.]

The outline analysis of variance is as follows.

SOURCE	DF	
Blocks	7	[Note that ABC, ADE, BCDE are included here]
Main effects (A to E)	5	
2-factor interactions	10	
3-factor interactions	8	[Excluding ABC and ADE]
4-factor interactions	4	[Excluding ABDE]
5-factor interaction ABCDE	1	
Residual	28	
TOTAL	63	

Solution continued on next page

(ii) Now there are only enough units for a single complete replicate. Again they need grouping into blocks of 8.

To conduct a full 2^6 experiment in eight blocks of size 2^3, three contrasts must be selected for confounding; their "products" will also be confounded. A scheme for doing this that only confounds 3- and 4-factor interactions is to select three 3-factor interactions for which any pair have only one letter in common, such as ABC, CDE and ADF. The full confounding scheme is given as follows, with the "generators" underlined:

I <u>ABC</u> <u>CDE</u> ABDE <u>ADF</u> BCDF ACEF BEF.

The outline analysis of variance is as follows.

SOURCE	DF	
Blocks	7	[Including all confounded interactions]
Main effects (A to F)	6	
2-factor interactions	15	
3-factor interactions	16	[Excluding ABC, CDE, ADF and BEF]
4-factor interactions	12	[Excluding ABDE, BCDF and ACEF]
5-factor interactions	6	
6-factor interact'n ABCDEF	1	
TOTAL	63	

There is no residual as such as there is no replication in this design.

(iii) Analysis of the results from the experiment in part (ii) requires use of some of the interactions as residual. Although this can be unsatisfactory, high-order interactions are unlikely to exist as such (and would be extremely difficult to interpret if they did exist), and usually there would be little or no hesitation in using 6- and 5-factor interactions as residual, thus already giving 7 d.f. for this purpose. Further, it might well be acceptable to add in the unconfounded 4-factor interactions, giving 19 d.f. for "residual". All the main effects and 2-factor interactions can be identified in the analysis, and so can most of the 3-factor interactions.

This gives a reasonably satisfactory analysis for the experiment in part (ii) and so, as the sixth factor (F) is thought to be of importance, this would be preferred to the experiment in part (i).

Graduate Diploma, Module 4, 2010. Question 3

[Solution continues on next two pages]

(a) [Summations in this solution are all from 1 to v.]

Given a set of coefficients $\{c_i\}$ such that $\Sigma c_i = 0$, the random variable $\Sigma c_i T_i$ is a linear contrast between the treatment totals $\{T_i\}$.

Now letting t_i denote the observed value for each T_i, the value of the contrast is $\Sigma c_i t_i$ and the sum of squares for it is $\{\Sigma c_i t_i\}^2 / r\Sigma c_i^2$ [note: r is the number of replicates, as given in the question].

Suppose that $\Sigma d_i T_i$ is a second contrast between the $\{T_i\}$, so that $\Sigma d_i = 0$. If $\Sigma c_i d_i = 0$, these two contrasts are orthogonal.

Orthogonal contrasts lead to independent terms in an analysis of variance and thus to independent tests and comparisons. Further, the sums of squares for a complete set of orthogonal contrasts add up to the total treatment sum of squares in the analysis of variance and thus, by testing each one against the residual in the usual way, give a mechanism for investigating where any treatment differences lie.

(b) (i) It is sensible to examine whether there is any overall difference at all between the untreated control O and the treatments: "O versus rest". This can be examined by adding the totals for all the treatments and subtracting this sum from five times the total for O; this would be expected to be zero in the absence of any overall difference, but non-zero if there is an overall difference. A contrast for doing this may be written as follows.

Treatments	O	S	A	B	C	D
Coefficients	5	–1	–1	–1	–1	–1

It is also sensible to examine whether there is any overall difference between the standard treatment S and the other, experimental, treatments: "S versus A, B, C, D". In a similar way, a contrast for doing this may be written as follows.

Treatments	O	S	A	B	C	D
Coefficients	0	4	–1	–1	–1	–1

An overall comparison between (A, B) and (C, D) would also be of interest. A contrast for doing this may be written as follows.

Treatments	O	S	A	B	C	D
Coefficients	0	0	1	1	–1	–1

A comparison between A and B would be of interest. A contrast for doing this may be written as follows.

Treatments	O	S	A	B	C	D
Coefficients	0	0	1	–1	0	0

Similarly, a contrast for comparing C and D may be written as follows.

Treatments	O	S	A	B	C	D
Coefficients	0	0	0	0	1	–1

The full set of five contrasts is as follows.

Treatments	O	S	A	B	C	D
O versus rests	5	–1	–1	–1	–1	–1
S versus A,B,C,D	0	4	–1	–1	–1	–1
A,B versus C,D	0	0	1	1	–1	–1
A versus B	0	0	1	–1	0	0
C versus D	0	0	0	0	1	–1

Inspection of these contrasts readily shows that they are orthogonal. In the context of the problem at hand, this is because they were created to make independent comparisons.

(ii) We need the values and thus the sums of squares for these contrasts. To illustrate the method, the value for the "O versus rest" contrast is

$$(5 \times 45) + \{(-1) \times 55\} + \{(-1) \times 70\} + \{(-1) \times 72\} + \{(-1) \times 63\} + \{(-1) \times 87\}$$
$$= -122$$

and thus the sum of squares is $(-122)^2/(5 \times 30) = 99.227$.

The complete analysis of variance is as follows. The overall treatment sum of squares (212.267, given in the question, with 5 d.f.) has been split up into the separate components identified by the contrasts.

SOURCE	DF	SS	MS	*F* value
O vs rest	1	99.227	99.227	10.80 highly significant
S vs A,B,C, D	1	51.840	51.840	5.64 significant
A,B vs C,D	1	3.200	3.200	0.35 not significant
A vs B	1	0.400	0.400	0.04 not significant
C vs D	1	57.600	57.600	6.27 significant
Residual	24	220.533	9.189	$= \hat{\sigma}^2$
TOTAL	29	432.800		

[All *F* values are compared with $F_{1,24}$; upper 5% point is 4.26, upper 1% point is 7.82, upper 0.1% point is 14.03.]

There is strong evidence that there is a real overall difference between the untreated control O and the rest of the treatments. The totals clearly indicate that all the other treatments give higher mean readings than O.

There is evidence that there is a real overall difference between the standard treatment S and the experimental treatments A, B, C and D. The totals suggest that all the experimental treatments give higher mean readings than S, though there may be some doubt as to whether this is a real difference in the case of experimental treatment C.

There is no evidence of an overall average difference between the effects of the two compounds, one of which is represented by treatments A and B and the other by treatments C and D.

There is no evidence of an overall average difference between the effects of treatments A and B – that is, the physical forms of the compound used in these treatments do not appear to lead to different results.

There is evidence of an overall average difference between the effects of treatments C and D – that is, the strengths at which this compound is used do appear to lead to different results, with strength C giving lower readings than strength D. This facet of these treatments complicates some of the other interpretations.

Graduate Diploma, Module 4, 2010. Question 4

(i) Variety (V) totals are: 155, 139, 140. These are totals of 8 observations.

Level (L) totals are: 87, 100, 126, 121. These are totals of 6 observations.

Treatment totals (of 2 observations) are given in the question.

The grand total (of 24 observations) is 434.

"Correction term" is $\frac{434^2}{24} = 7848.167$.

Therefore total SS = 8202 – 7848.167 = 353.833.

Overall SS for treatments

$$= \frac{34^2}{2} + \frac{39^2}{2} + ... + \frac{48^2}{2} - 7848.167 = 8149.000 - 7848.167 = 300.833.$$

SS for varieties (V)

$$= \frac{155^2}{8} + \frac{139^2}{8} + \frac{140^2}{8} - 7848.167 = 7868.250 - 7848.167 = 20.083.$$

SS for levels (L)

$$= \frac{87^2}{6} + .. + \frac{121^2}{6} - 7848.167 = 8014.333 - 7848.167 = 166.167.$$

Therefore by subtraction the SS for the VL interaction is

300.833 – 20.083 – 166.167 = 114.583.

Also by subtraction the residual SS is 353.833 – 300.833 = 53.000.

The numbers of degrees of freedom are

23 for the total SS

11 for the overall treatments SS, consisting of

- 2 for V
- 3 for L
- 6 (= 2 × 3) for the VL interaction

12 for the residual SS

Solution continued on next page

Hence:

SOURCE	DF	SS	MS	F value
Varieties (V)	2	20.083	10.042	2.27
Levels (L)	3	166.167	55.389	12.54
VL interaction	6	114.583	19.097	4.32
Treatments	11	300.833	–	–
Residual	12	53.000	4.417	$= \hat{\sigma}^2$
TOTAL	24	353.833		

The F value of 2.27 is referred to $F_{2,12}$; this is not significant (the upper 5% point is 3.89), so there is no evidence of overall differences among the varieties.

The F value of 12.54 is referred to $F_{3,12}$; this is beyond the upper 0.1% point (which is 10.80), so there is extremely strong evidence of an effect of the level of soil nutrient.

The F value of 4.32 is referred to $F_{6,12}$; this is beyond the upper 5% point (3.00) but not beyond the upper 1% point (4.82), so there is some evidence of an interactive effect between variety and level.

The "value" for the linear component of L is $(-3\times87) - 100 + 126 + (3 \times 121) = 128$.

As each of the totals feeding into this calculation is a total of 6 observations, the divisor for the SS for this contrast is $6 \times \{(-3)^2 + (-1)^2 + 1^2 + 3^2\} = 120$, and thus the SS for the linear component of L is $128^2/120 = 136.533$, with one degree of freedom.

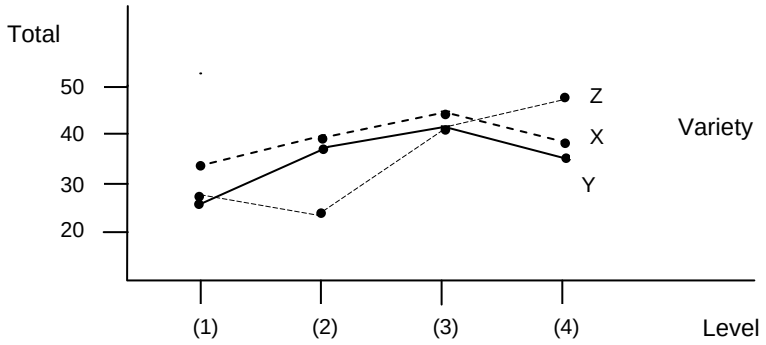
Comparing this with the residual MS in the usual way gives an F value of 136.533/4.417 = 30.91, which is referred to $F_{1,12}$. This is well beyond the upper 0.1% point (which is 18.64), so there is extremely strong evidence of a linear component in the L effect.

The SS for the remainder of the L effect is 166.167 – 136.533 = 29.634, with two degrees of freedom. This is clearly very small compared with the linear component, and can be formally tested in the usual way: its F value is (29.634/2)/4.417 = 3.35 which is not significant when referred to $F_{2,12}$ (upper 5% point is 3.89).

So we may conclude that the scientist is justified in believing that there is a linear relation with the quantity of soil nutrient – this seems sound in outline.

Solution continued on next page

(ii)



Variety Z differs from X and Y in having a "dip" at level (2) but continuing to rise through level (3) to level (4). Varieties X and Y seem to show a linear increase up to level (3) but then drop at level (4). These features are illustrations of the interaction between varieties and levels.

(iii) The scientist needs to study the possible non-linearity for variety Z at the lower levels of soil nutrient and likewise the possible non-linearity for varieties X and Y at the upper levels.

Graduate Diploma, Module 4, 2010. Question 5

Part (i)

The least squares estimator is given by $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$.

These estimators are linear functions of the observations **Y** each of which is unbiased for estimating its respective parameter. The Gauss-Markov theorem states that these estimators have minimum variance among all such linear unbiased estimators.

An example that is not a linear model is $Y = Axe^{\lambda x} + \varepsilon$. (Note that this model is not linear in the parameter (λ).)

Part (ii)

(a) The models are all *linear in the parameters*.

(b) This is model B. This can be written out as $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ (+ error) where x_1 represents the temperature and each of x_2 and x_3 takes value 0 or 1 to define the alloy. Thus we obtain parallel straight lines with a different intercept for each alloy.

(c) In model A, the electrical resistivity depends only on the temperature (x_1). So this model represents a single straight line relationship between electrical resistivity and temperature for all three alloys.

In model C, there are three straight lines with the same intercept but different slopes.

In model D, there are three straight lines with different slopes and different intercepts.

(d) Model A is clearly very poor – its regression sum of squares is very small.

To choose between the others, we start with the simplest which is model B – same slopes for the three alloys but different intercepts. This model has 3 parameters, so its regression sum of squares (35232) has 3 degrees of freedom. Including different slopes (i.e. moving to model D) gives a regression sum of squares of 35517 with 5 degrees of freedom (as there are 5 parameters in this model).

The total sum of squares is 35673 and this has 26 degrees of freedom (note that there are 27 observations). So we can construct an analysis of variance based on the Extra Sum of Squares Principle as follows.

Solution continued on next page

SOURCE	DF	SS	MS	F value
Intercepts only (model B)	3	35232	–	–
Including slopes	2	285	142.5	19.18
Intercepts and slopes (model D)	5	35517	–	–
Residual	21	156	7.43	$= \hat{\sigma}^2$
TOTAL	26	35673		

The F value of 19.18 is referred to $F_{2,21}$; this is extremely highly significant (the upper 0.1% point is between 9.95 and 9.61, respectively the upper 0.1% points for $F_{2,20}$ and $F_{2,22}$), so there is extremely strong evidence that model D is an improvement over model B.

(A similar formal analysis could be used to show that model D is also superior to model C, but it is scarcely necessary to consider this as model C is more complicated than model B but has a smaller regression sum of squares.)

(e) For model D, R^2 = 35517/35673 = 0.9956 (i.e. 99.56%), so that part of the physicist's statement is correct. However, a high value of R^2 does not, in itself, guarantee a good fit and a correct model. Other diagnostics need to be considered too – for example, residual plots.

(f) Prior information from other studies may suggest other models that should be examined (but care should be taken that there are some substantive reasons why these models might be suitable, otherwise dangers of excessive "data dredging" may obtrude).

Plots of resistivity against temperature for the different alloys may show that curvature is present.

Residual plots can suggest particular types of lack of fit.

It may be useful to obtain more observations if curvature is to be studied seriously.

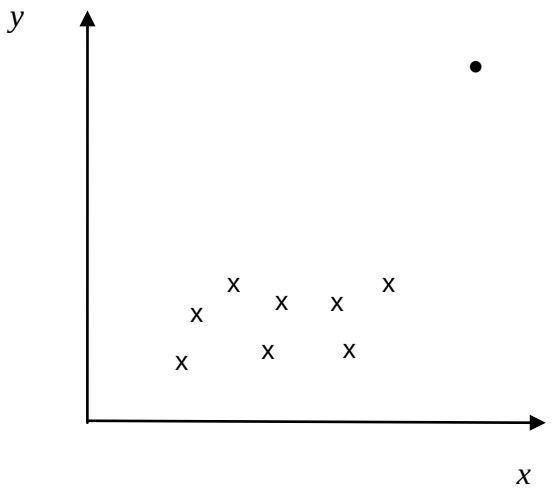
Graduate Diploma, Module 4, 2010. Question 6

An observation has high leverage if it is an outlier and it has a disproportionate effect on the form of the model.

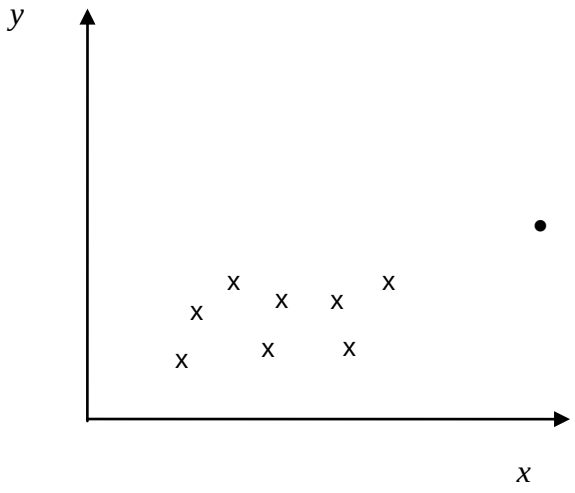
An observation is influential if its omission would cause an important change to the model.

The leverage values are the diagonal elements of the "hat" matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$.

(i) Observation with high leverage and high influence.

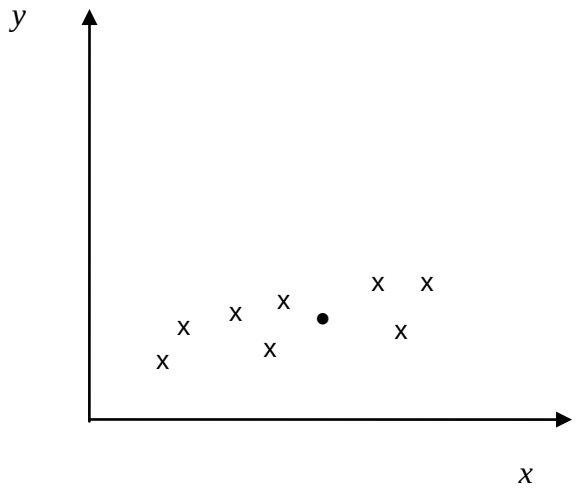


(ii) Observation with high leverage and low influence.



Solution continued on next page

(iii) Observation with low leverage and low influence.



Each standardised residual r_i is given by $r_i = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ where $\hat{\varepsilon}_i$ is the residual, $\hat{\sigma}$ is the required estimate of the residual standard deviation and h_{ii} is the (i, i) element of the hat matrix, i.e. it is the leverage value.

Using case 3 gives $1.62 = \dfrac{2680}{\hat{\sigma}\sqrt{1-0.60}}$ so that $\hat{\sigma}$ = 2616. (It may be confirmed that similar values are obtained from the other cases.)

Thus the three missing leverage values are

case 1 $\quad \sqrt{(1 - h_{11})} = 4503/(2616 \times 2.39) = 0.7202, \quad \therefore\ h_{11} = 0.48$

case 2 $\quad \sqrt{(1 - h_{22})} = 1604/(2616 \times 1.06) = 0.5784, \quad \therefore\ h_{22} = 0.67$

case 4 $\quad \sqrt{(1 - h_{44})} = -889/(2616 \times (-0.47)) = 0.7230, \quad \therefore\ h_{44} = 0.48.$

Cook's distance is a measure of influence. It is high for cases 1 and 3. Leverage values in excess of 0.5 are often considered as indicating observations with high leverage. Here, cases 2, 3, 7, 10 and 12 all come into that category; and there are several others with leverage values only slightly less than 0.5. All this indicates that the model used is a poor fit to the data.

Many of the residuals are large (in absolute value) compared with the observed or fitted values; and, more importantly in view of the information that the cases are numbered in the order of being observed, the first three residuals are positive and all the rest are negative. This suggests there may be a missing term that would model curvature. We also note that case 1 is by a very long way the highest observed cost, and case 2 is the lowest. It may be difficult to model this initial behaviour.

It seems that, at the least, a more general second-order model with squares and cross-products should be examined.

Graduate Diploma, Module 4, 2010. Question 7

Part (i)

$(SM)_{ij}$ is an interaction term which measures whether or not the different sexes show the same job satisfaction in otherwise identical circumstances.

Similarly, $(SMA)_{ijk}$ measures whether job satisfaction is different for different sexes by a different amount in different areas.

Part (ii)

(a) In this dataset, the numbers of males and females are the same. This means that the coefficient for sex must be zero in a first-order model. However, it must be retained in the model as it will be needed if interactions are also included.

(b) Let *LP* denote the linear predictor for satisfied males in area 1, with subscript 1 for the first model (i.e. the model shown in the computer output in the question in the preamble of part (ii)) and subscript 2 for the second model (i.e. that shown in part (ii)(b) in the question).

The value of LP_1 is given by *LP* = constant + "satisfied" + "area 1" + "male".

So we have

$$LP_1 = 4.5010 + 0.1939 + 0 + 0 = 4.6949,$$
$$LP_2 = 4.2551 + 0.0970 + 0.3428 + 0 = 4.6949.$$

The two analysts have merely used different codings for *S* and *A*. Either model can be written in the same form as the other. To illustrate, the first model can be written in the form of the second as

Factor	*Level*	*Coefficient*
Constant		4.5010
S	Satisfied	0.1939
	Not satisfied	0.0000
M	Male	0.0000
	Female	0.0000
A	Area 1	0.0000
	Area 2	–0.6857

and we see that (within the limits of rounding) $0.1939 = 2 \times 0.0970$, $-0.6857 = 2 \times (-0.3428)$ and $4.5010 = 4.2551 - 0.0970 + 0.3428$. Thus the models are equivalent.

Solution continued on next page

Part (iii)

(a) The scaled deviance for the full model is 0.

(b) Backward elimination begins with consideration of the smallest change from the full model, which is to omit *SMA* as this gives a scaled deviance of 0.53 which is the smallest of the scaled deviances. The change (0.53 – 0) is not significant as an observation from χ^2 with one degree of freedom, so this is not a significant change and we adopt the model with *SMA* omitted.

The smallest change now is to omit *SA*, giving a scaled deviance of 2.99. The change is 2.99 – 0.53 = 2.46 which is again not significant as an observation from χ^2 with one degree of freedom. So this also is not a significant change, and we adopt the model with *SA* also omitted.

To consider reduction of the model still further, we investigate the removal also of *SM* as this gives the smallest change, to the model with scaled deviance 8.66. The change is 8.66 – 2.99 = 5.67, and this is a significant change – so we must keep *SM* in the model.

Thus the model arrived at by backward elimination contains μ, *S*, *M*, *A*, *SM* and *AM*.

This model can be confirmed as a reasonable fit by considering its scaled deviance divided by its number of degrees of freedom (6, as there are 6 parameters in the model): 2.99/6 is less than 1, which indicates, somewhat informally, a reasonable fit.

(c) We need to consider the interactions with satisfaction, not the main effects. Adding *SA* to the main effects makes very little difference to the scaled deviance, so it appears that area (*A*) does not affect satisfaction. However, adding *SM* to the main effects gives a significant reduction in scaled deviance. So it appears that sex (*M*) does affect job satisfaction.

Graduate Diploma, Module 4, 2010. Question 8

(i) The logistic regression model for π, the probability of being prepared to take part, is

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where the logit transformation is $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$.

(ii) Using 0, 1, 2 to code age would assume that there is the same difference in results between the first and second age groups as between the second and third. It may be unwise to assume in advance that this is true – and the coefficients of the fitted model do not support it. On the other hand, the current coding with x_1 and x_2 uses up an extra degree of freedom.

On the whole, in this case it is probably better to use the current coding.

(iii) (a) In the given output, all the standard errors are large compared with the estimates of the coefficients. So we can conclude that the pilot study has found that none of these coefficients is significantly different from zero. (Of course this may not be the case in the full survey that will follow later – this is touched on in part (iii)(c).)

(b) An approximate 95% confidence interval for the x_1 coefficient is

$$0.0884 \pm (1.96 \times 0.2532) = 0.0884 \pm 0.4963,$$

i.e. it is (–0.4078, 0.5847).

(c) As a pilot study, it has been kept fairly small, and so it is likely to be under-powered. More useful information will be found from the widths of the confidence intervals for each coefficient than by merely testing whether coefficients are significantly different from zero (which is of course equivalent to checking whether zero is in the interval) – though the intervals will be wide.

Solution continued on next page

(iv) For females over 59, we have $x_1 = 1$, $x_2 = 0$, $x_3 = 1$. Inserting these in the fitted model gives

$$\log\left(\frac{\pi}{1-\pi}\right) = 0.0655 + 0.0884 - 0.1953 = -0.0414\,,$$

so $\dfrac{\pi}{1-\pi} = e^{-0.0414}$ and so the estimate of π is $\dfrac{e^{-0.0414}}{1+e^{-0.0414}} = 0.4897$.

Thus, based on data from the pilot study, about 49% of females over 59 would be prepared to take part in the survey.

(v) 0.0884 is the coefficient for sex and relates to females ($x_1 = 1$).

Consider two people in the same age group but of different sexes. The fitted model gives the following.

For a female: $$\log\left(\frac{\pi_F}{1-\pi_F}\right) = \text{constant} + 0.0884 + (x_2 \text{ term}) + (x_3 \text{ term})$$

For a male: $$\log\left(\frac{\pi_M}{1-\pi_M}\right) = \text{constant} + 0 + (x_2 \text{ term}) + (x_3 \text{ term})$$

$$\therefore\ \log\left(\frac{\pi_F}{1-\pi_F}\right) - \log\left(\frac{\pi_M}{1-\pi_M}\right) = 0.0884\,.$$

$$\therefore\ \frac{\dfrac{\pi_F}{1-\pi_F}}{\dfrac{\pi_M}{1-\pi_M}} = e^{0.0884} = 1.09\,.$$

This is the odds ratio for females relative to males, i.e. $e^{\text{coefficient}}$ = odds ratio. It shows that, according to the pilot study, females are very marginally more likely than males within the same age group to be prepared to take part in the survey.