THE ROYAL STATISTICAL SOCIETY

2009 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

MODULAR FORMAT

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, 2009. Question 1

(i) The response variable y is a continuous random variable. Each predictor variable x_i is either continuous or binary, and regarded as fixed without experimental error. The form of the model is valid with no important predictor omitted. The residual (error) term ε covers all random departures from the model, and has $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$ (a constant), and the $\{\varepsilon\}$ are a mutually uncorrelated set; for analysis, Normality is also usually assumed.

(ii) The matrix formulation is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the column vector of the values of y (i.e. the observations), $\boldsymbol{\beta}$ is the column vector of the parameters β_i, $\boldsymbol{\varepsilon}$ is the column vector of the residuals ε and $\mathbf{X}$ is the matrix (often called the design matrix) given as follows, where there are n observations on each variable:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & & & & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}.$$

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

(iii) (a) The estimators are linear functions of the observations y. They are unbiased estimators of their respective parameters, having the minimum variance among all such linear unbiased estimators. Their variances and covariances are given by the variance matrix

$$\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

(b) The estimators are Normally distributed (and hence confidence intervals and t and F tests involving parameters in the model can be obtained). The estimators are also the maximum likelihood estimators.

(iv) Highly dependent predictor variables may lead to high correlations among the $\{x_i\}$ and this can make the $\mathbf{X}^T\mathbf{X}$ matrix singular or nearly so. If it is actually singular, it cannot be inverted and so the estimators will not exist. If it is nearly singular, it is likely that $\mathbf{X}^T\mathbf{X}$ will be computationally unstable, and the variances of the estimators may be very large; thus the estimators are unreliable.

Solution continued on next page

Possible methods to overcome such problems are

to use practical knowledge, perhaps from previous experience, to remove redundant variables and then carry out the analysis using those of most practical importance

to use a data reduction method such as principal component analysis to reduce the number of variables, identifying their best combinations

to use robust or ridge regression methods.

(v) R^2 takes no account of the number of variables in the model, being simply a measure of the amount of the total variation about the mean that is accounted for by the fitted model. If more variables are added to the model, R^2 must increase (or possibly remain unaltered, but never decrease).

Some workers therefore prefer to adjust R^2 to take account of the number of variables in the model, by using

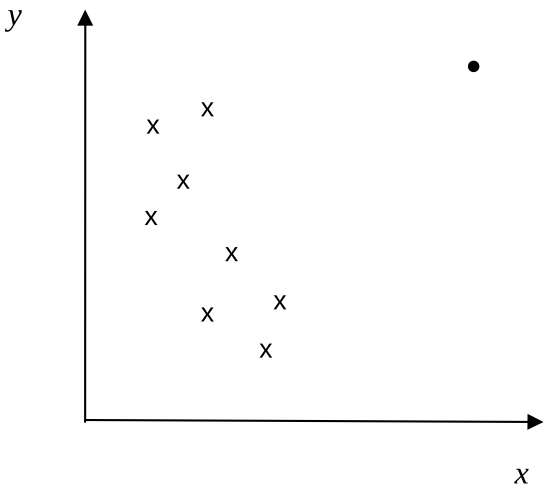
$$\text{adjusted } R^2 = 1-\left(1-R^2\right)\left(\frac{n-1}{n-p}\right).$$

(vi) In some situations, the parameter estimates may change substantially if one particular case of data (y_i, x_{1i}, x_{2i}, ..., x_{pi}) is omitted from the analysis. Such a case is called influential. The conclusions from the full analysis and from that with the influential point omitted may both be unreliable, perhaps especially so if there are no other data points near to the influential one. An example where there is just one x variable is sketched in the diagram on the next page, • being the influential point.

Diagnostics obtained from statistical packages compare the two vectors $\hat{\boldsymbol{\beta}}$ with and without the influential point. One method is Cook's distance, $D_i = \left(\hat{\mathbf{Y}}-\hat{\mathbf{Y}}_{(i)}\right)^T\left(\hat{\mathbf{Y}}-\hat{\mathbf{Y}}_{(i)}\right)/\,p\hat{\sigma}^2$, where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is the vector of fitted values obtained when all data points are used in the calculation and $\hat{\mathbf{Y}}_{(i)}$ is that when one point (i) is not used, $\hat{\sigma}^2$ is the residual mean square for the fitted model (using all data points) and p is the number of fitted parameters. A "large" value (commonly taken as greater than 1) of D_i indicates that the ith point may be influential and merits further examination. DFFITS works in a similar way but using the residual mean square without the ith item.

Leverage values are also often quoted; these are based on the x-values rather than on the observations y.

Solution continued on next page


y
x
x
x
x
x
x
x
x
x
x

Graduate Diploma, Module 4, 2009. Question 2

Part (i)

When the blocks in an experiment are not sufficiently large to accommodate all the treatments, it is obviously not possible to ensure that each treatment appears in each block. This destroys the symmetry of the usual (complete) randomised blocks design. A balanced incomplete block design preserves a sufficient degree of symmetry to enable all pairs of treatments to be compared with the same precision. Thus, in terms of the usual notation, a balanced incomplete block design is useful when the block size k is smaller than the number of treatments v and all comparisons between pairs of treatments are regarded as equally important.

The total number of experimental units ("plots") is equal to the number of treatments (v) multiplied by the number of replicates of each treatment (r), and also equal to the number of blocks (b) multiplied by the number of units in each block (k). Thus we have

$$rv = bk.$$

Now consider any particular treatment. Over the entire design, it appears in r blocks, in each of which there are also $k - 1$ other treatments. This gives a total of $r(k - 1)$ other experimental units which are in blocks that contain a unit with this treatment. These units have to contain the other $v - 1$ treatments λ times each. So we have

$$\lambda(v - 1) = r(k - 1).$$

Note that λ is an integer, so $r(k - 1)/(v - 1)$ must be an integer. A balanced incomplete block design only exists if the values of r, k and v are such that this is so.

Part (ii)

(a) $v = 5$ (A, B, C, D, E) $b = 5$ (batches 1 to 5) $r = 4$ $k = 4$ $\lambda = 3$

(b) The total SS is $955360 - \frac{4348^2}{20} = 10104.80$.

The total for batch 1 (block 1) is 194 + 205 + 250 + 214 = 863 and similarly for the other blocks, so the SS for batches (blocks) is

$$\frac{863^2}{4} + \frac{883^2}{4} + \frac{835^2}{4} + \frac{861^2}{4} + \frac{906^2}{4} - \frac{4348^2}{20} = 704.80\,.$$

The SS for treatments adjusted for batches is [formula quoted in question]

$$\frac{1}{60}\left((-421)^2 + (-142)^2 + 311^2 + (-215)^2 + 467^2\right) = \frac{558440}{60} = 9307.33\,.$$

Solution continued on next page

Hence:

SOURCE	DF	SS	MS	*F* value
Batches	4	704.80	–	
Treatments adjusted for batches	4	9307.33	2326.83	276.2
Residual	11	92.67	8.425	$= \hat{\sigma}^2$
TOTAL	19	10104.80		

The *F* value of 276.2 is referred to $F_{4,11}$; this is very highly significant indeed (the upper 0.1% point is 10.35), so there is overwhelming evidence that there are differences between the treatments, having adjusted for the blocks. The differences are explored in part (c).

(c) The adjusted treatment means [formula quoted in question] are as follows.

For A: $\frac{4348}{20} + \frac{-421}{15} = 189.33$

For B: $\frac{4348}{20} + \frac{-142}{15} = 207.93$

For C: $\frac{4348}{20} + \frac{311}{15} = 238.13$

For D: $\frac{4348}{20} + \frac{-215}{15} = 203.07$

For E: $\frac{4348}{20} + \frac{467}{15} = 248.53$

The table below shows these in ascending order.

A	D	B	C	E
189.33	203.07	207.93	238.13	248.53

The variance of the difference between any pair of treatment means is estimated by [formula for variance is quoted in the question]

$$\frac{2k\hat{\sigma}^2}{v\lambda} = \frac{2 \times 4 \times 8.425}{5 \times 3} = 4.493 .$$

Least significant differences are therefore given by $t \times \sqrt{4.493}$ where *t* denotes the appropriate critical point from the t_{11} distribution: 2.201 for 5%, 3.106 for 1%, 4.437 for 0.1%. So the respective LSDs are 4.67, 6.58, 9.41.

Solution continued on next page

A and D represent 10% cadmium without and with tin. The difference is highly significant – there is strong evidence that introducing tin (at 10%) when using 10% cadmium gives increased melting point.

C and E represent 30% cadmium without and with tin. This difference is also highly significant – there is strong evidence that introducing tin (at 10%) when using 30% cadmium gives increased melting point.

The differences (A, B) and (B, C) represent increasing percentages of cadmium with no tin. These differences are highly significant. There is strong evidence that increasing cadmium from 10% to 20% and to 30% in the absence of tin gives increased melting point.

In summary, each increase in cadmium or in tin appears to increase the melting point.

Graduate Diploma, Module 4, 2009. Question 3

(i) The eight treatment combinations used all have an odd number of letters in common with ACE and with BDE. Hence I = ACE and I = BDE form the defining contrast, together with their generalised interaction ABCD.

(ii) We have

I = ACE = BDE = ABCD

which gives

A = CE = ABDE = BCD
B = ABCE = DE = ACD
C = AE = BCDE = ABD
D = ACDE = BE = ABD
E = AC = BD = ABCDE
AB = BCE = ADE = CD
AD = CDE = ADE = BC.

(iii) In the table below, the coefficients +1 or –1 are indicated by + or –.

"Value" (sometimes referred to as "total effect") is calculated using the coefficients; eg for A, "value" = –8.7 + 12.0 – 17.5 + 11.0 – ... + 17.7.

"Estimate", the estimate of the main effect or interaction, is simply "value"/4.

All the remaining interactions are aliased.

	e	*ad*	*bde*	*ab*	*cd*	*ace*	*bc*	*abcde*	Value	Estimate
	8.7	12.0	17.5	11.0	9.0	13.0	16.1	17.7		
A	–	+	–	+	–	+	–	+	2.4	0.60
B	–	–	+	+	–	–	+	+	19.6	4.90
C	–	–	–	–	+	+	+	+	6.6	1.65
D	–	+	+	–	+	–	–	+	7.4	1.85
E	+	–	+	–	–	+	–	+	8.8	2.20
AB	+	–	–	+	+	–	–	+	–12.2	–3.05
AD	+	+	–	–	–	–	+	+	4.0	1.00

(iv) The two-way table of means for AB is calculated thus:

	B^-		B^+	
A^-	*e*, *cd*	8.85	*bc*, *bde*	16.80
A^+	*ad*, *ace*	12.50	*ab*, *abcde*	14.35

Because of the alias structure, this could equally well be the two-way table for CD.

Solution continued on next page

(v) In general, fractional factorial designs should at least allow all main effects and two-factor interactions to be estimated (but note that it has not been possible to achieve this in the present case – there are several cases of main effects aliased with two-factor interactions and of two-factor interactions aliased with each other; for 5 factors, blocks of size 8 are not really large enough, a block size of at least 16 is desirable), unless some have already been studied in earlier experiments. Higher-order interactions may be used for "residual" *if there is some reason* to suppose that they are negligible, from previous experiments or from other knowledge about the factors. It is *not* valid to construct a "residual" merely by inspection of the analysis of the data – an interaction that appears small may indeed be a real but small interaction, rather than a manifestation of experimental error.

In the present case, all terms in the analysis represent main effects and/or two-factor interactions. No matter how "small" some of these may appear to be, it would be wrong to use any of them as "residual" (unless it was already known that some are in fact negligible – but, if so, one might question the wisdom of designing the present experiment in this manner anyway).

Sometimes designs of this nature are used in response surface analysis, in which case the aim is fitting models rather than estimation and different criteria may apply.

[In the examination, suitable credit, up to a maximum of the stated tariff of 4 marks, was given for any relevant comments.]

Part (i)

In a fixed effects model, the parameters are constants to be estimated, for example for treatment effects and block effects in a randomised blocks design, and inferences will only apply to the actual set of treatments included in the experiment (and the conditions under which it was done).

In a random effects model, the parameters are random variables, the particular levels used in the experiment being a random sample from a much wider population of levels which could have been used, for example all the (many) hospitals in a region from which only a few are selected (at random) for study. Inferences are assumed to be valid for this wider population. Variability is the main item for study, in terms of the variance of the wider population.

Part (ii)

(a) y_{ijk} is the tensile strength measured on the kth beam (k = 1, 2, 3) in the jth batch (j = 1, 2, 3, 4) within the ith site (i = 1, 2, 3).

μ is a fixed term representing the overall grand mean.

Since the company has only three sites, with no suggestion that these are a sample from a wider population of sites, s_i is a fixed term representing site effects, with $\Sigma s_i = 0$.

The batches have been selected at random within each site, so b_{ij} is a random term representing variation between batches at site i. It is assumed that $b_{ij} \sim N(0, \sigma_b^2)$. Note that σ_b^2 is assumed to be constant over i, i.e. the underlying variance is the same for each site.

ε_{ijk} is a random residual (error) term, distributed as $N(0, \sigma^2)$ for all items.

(b) There are 36 observations and so 35 degrees of freedom in total. There are 3 sites, so 2 degrees of freedom for sites. Each site has four batches giving 3 degrees of freedom, and therefore there are 9 degrees of freedom altogether for variation between batches within sites. This leaves 24 degrees of freedom for variation within batches (this could be thought of as 2 degrees of freedom for each of the 12 batches).

The completed analysis of variance is therefore as shown below.

Solution continued on next page

Source of variation	d.f.	Sum of squares	Mean square	*F* value
Between sites	2	418.72	209.3600	209.3600/189.0656 = 1.11
Between batches within sites	9	1701.59	189.0656	189.0656/10.4167 = 18.15
Within batches	24	250.00	10.4167	
Total	35			

To test the null hypothesis "all $s_i = 0$", 1.11 is referred to the $F_{2,9}$ distribution. This is not significant at even the 10% significance level (the upper 10% point is 3.01), so there is no evidence against this null hypothesis. We may assume that there are no differences between the means for the sites.

To test the null hypothesis that $\sigma_b^2 = 0$, 18.15 is referred to the $F_{9,24}$ distribution. This is very highly significant – the upper 0.1% point is 4.80 – so we reject this null hypothesis. There is very strong evidence that there is variability between the batches within the sites.

σ^2, the variance within batches, is estimated by 10.4167.

σ_b^2 is estimated by $\dfrac{189.0656-10.4167}{3} = 59.55$ [see the expected mean squares given in the question]. Note that this is about six times the estimated variance within batches.

In summary, it appears that there is no variability between sites but an important source of variability between batches within sites.

Graduate Diploma, Module 4, 2009. Question 5

(i) $y_{ijk} = \mu + S_i + E_j + (SE)_{ij} + \varepsilon_{ijk}$ $\quad i = 1, 2, 3; \quad j = 1, 2, 3; \quad k = 1, 2, 3.$

y_{ijk} is the kth observation (fuel consumption, miles per gallon (mpg)) at speed i with engine size j

μ is the overall grand mean

S_i is a speed effect (departure from μ) for speed i, with $\Sigma S_i = 0$

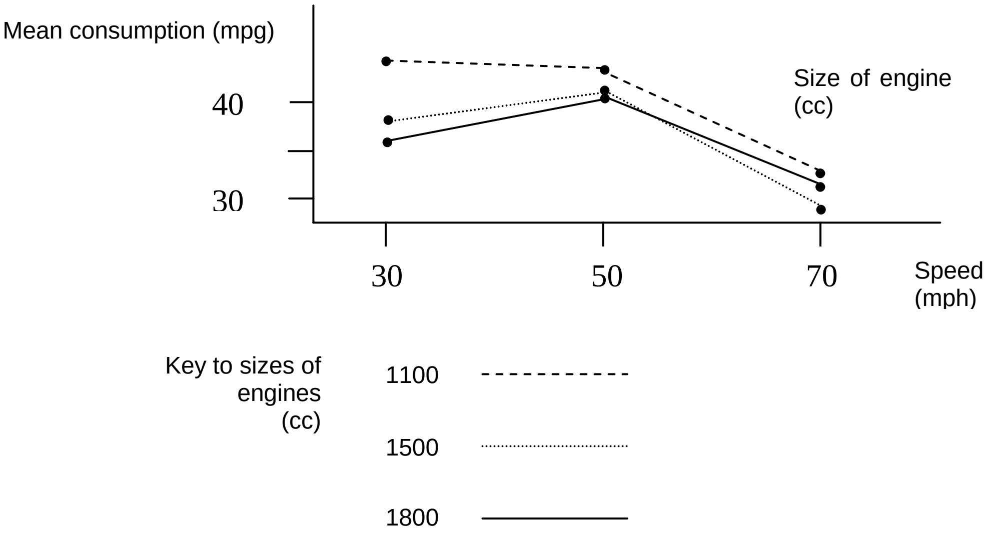
E_j is an engine size effect (departure from μ) for engine size j, with $\Sigma E_j = 0$

$(SE)_{ij}$ is an interaction between S_i and E_j, with $\Sigma_i(SE)_{ij} = 0$ and $\Sigma_j(SE)_{ij} = 0$

ε_{ijk} is a random residual (error) term, distributed $N(0, \sigma^2)$ [note constant σ^2 for all i, j and k]

(ii) The means are

Speed (mph)	Size of engine (cc)		
	1100	1500	1800
30	44.63	38.00	35.93
50	42.83	41.07	40.87
70	32.13	28.97	30.93



Solution continued on next page

(iii) The grand total is 1006.1 so the "correction factor" is $1006.1^2/27 = 37490.267$.

So the total sum of squares = 38253.85 – 37490.267 = 763.58, with 26 df.

SS for speeds = $\dfrac{355.7^2}{9}+\dfrac{374.3^2}{9}+\dfrac{276.1^2}{9}-37490.267 = 604.64$, with 2 df.

SS for engine sizes = $\dfrac{358.8^2}{9}+\dfrac{324.1^2}{9}+\dfrac{323.2^2}{9}-37490.267 = 91.57$, with 2 df.

Interaction SS = $\dfrac{133.9^2}{3}+\dfrac{114.0^2}{3}+\ldots+\dfrac{92.8^2}{3}-37940.267-604.64-91.57$

= 54.76, with 2×2 = 4 df.

The residual SS and df follow by subtraction.

Hence the analysis of variance is

Source of variation	d.f.	Sum of squares	Mean square	*F* value
Speeds (*S*)	2	604.64	302.32	431.5
Engine sizes (*E*)	2	91.57	45.79	65.4
S × *E* interaction	4	54.76	13.69	19.5
Residual	18	12.61	0.7006	$= \hat{\sigma}^2$
Total	26	763.58		

The *F* value of 431.5 is referred to $F_{2,18}$; this is well beyond the upper 0.1% point (which is 10.39), so there is extremely strong evidence of an effect of speed.

The *F* value of 65.4 is also referred to $F_{2,18}$ and is again well beyond the upper 0.1% point, so there is extremely strong evidence of an effect of engine size.

The *F* value of 19.5 is referred to $F_{4,18}$; this also is well beyond the upper 0.1% point (7.46 in this case), so there is extremely strong evidence of an interaction.

Solution continued on next page

(iv) Overall, though the effects of speed and engine size appear very highly significant, the results should be explained in terms of the interaction. This can be studied and explained both by reference to the diagram in part (ii) and more formally by considering least significant differences.

The variance of the difference between any pair of the means is estimated by

$$\frac{2}{3} \times 0.7006 = 0.4671 .$$

Least significant differences are therefore given by $t \times \sqrt{0.4671}$ where t denotes the appropriate critical point from the t_{18} distribution: 2.101 for 5%, 2.878 for 1%, 3.922 for 0.1%. So the respective LSDs are 1.43, 1.97, 2.68.

The diagram shows that, in the experiment, at any of the speeds the mean fuel consumption was better (higher mpg) for 1100 cc engines than for the other engine sizes. At 30 mph, this is a very highly significant difference – there is extremely strong evidence that the 1100 cc engines give a real improvement. At 50 mph the evidence is less compelling, but it still appears that there may be a real improvement. At 70 mph, the apparent improvement for 1100 cc engines compared with 1800 cc engines is within customarily accepted limits of experimental error, but there is still strong evidence that 1100 cc engines give a real improvement compared with 1500 cc engines at this speed.

The comparison between 1500 cc and 1800 cc engines is more complicated. At 30 mph, there is strong evidence that 1500 cc engines are better, but at 70 mph this situation is reversed. At 50 mph, there is no evidence at all for any difference in performances.

All three engine sizes show extremely strong evidence of a drop in fuel consumption performance from 50 mph to 70 mph; it appears that this is a substantial fall-off. There is likewise extremely strong evidence that the 1500 and 1800 cc engines give better performance at 50 mph than at 30 mph, whereas 1100 cc engines show some evidence of reduced performance at 50 mph compared with 30 mph.

Graduate Diploma, Module 4, 2009. Question 6

(i) In numerical computations, values in thousands (eg salaries in $) can lead to inaccuracies in sums of squares. More "manageable" values (eg after dividing by 1,000) may well be easier to use. The estimates and standard errors will be scaled by the same factor as the data, but can easily be scaled back to the original units. There will be no effect on the values of the usual test statistics or on their statistical significance. So the suggestion is sound.

(ii) The scatter plot shows the variability of current salary increasing as starting salary increases, ie it suggests that the underlying variance is non-constant. The "fan shape" suggests that a logarithmic transformation might validate the usual analysis more satisfactorily.

(iii) A model using log(current salary) as y and log(starting salary) as x would be of the form y = (other terms [sex, grade]) + $\beta_1 x$ + residual, which gives, on exponentiating,

$$\text{current salary} = \left(\text{expression for other terms}\right)\left(\text{starting salary}\right)^{\beta_1} \exp\left(\text{residual}\right)$$

which could make economic sense and be reasonably easy to interpret. On the other hand, using log(current salary) alone would lead to

log(current salary) = other terms + starting salary + residual

which is likely to be less easy to interpret.

The calculations would be simply carried out by a standard regression program with the variables transformed by taking logarithms.

(iv) To indicate the sex of the ith employee, define $s_i = 0$ or 1 according to whether the employee is male or female.

To indicate the grade of the ith employee, two dummy variables are required. For a clerk, define $t_i = m_i = 0$; for a team leader, define $t_i = 1$ and $m_i = 0$; for a manager, define $t_i = 0$ and $m_i = 1$.

Now let y_i = log(current salary) and x_i = log(starting salary), and let ε_i be the ith residual (error) term.

Then a suitable model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 s_i + \beta_3 t_i + \beta_4 m_i + \varepsilon_i \,.$$

Solution continued on next page

(v) There is no obvious pattern in the plot of standardised residuals against predicted values. This is good because it is indicative of a situation where a model fits the data equally well at all values of current salary.

There is likewise no obvious pattern in the plot of standardised residuals against logarithms of starting salary. This also is good; it suggests that using the logarithm of starting salary was sensible statistically, and there is no indication that some other function of starting salary should be considered for the modelling.

Most of the standardised residuals are between –2 and +2, which is what should be expected on the basis of an approximate N(0, 1) distribution for them. The number outside the interval (–2, 2) is not excessive for a data set of this size. However, most of those that are outside this interval are on the positive side of it, indicating possible slight skewness. It would be desirable to identify the observations that give rise to the values greater than 2, to see if there is some common characteristic (for example some special qualification possessed by these people).

There is one very large standardised residual, with a value close to +6. The observation giving rise to this should certainly be investigated individually.

(vi) (a) Using the dummy variables t_i and m_i defined in part (iv), the following additional terms could be added to the model: $\beta_{xt}x_i t_i$ and $\beta_{xm}x_i m_i$.

(b) Such an interaction would mean that the way in which starting salary influences current salary depends on the job grade of the employee.

Graduate Diploma, Module 4, 2009. Question 7

(i) Both forward selection and backward elimination are automatic methods based solely on statistical significance, ignoring any practical knowledge about the *x*-variables that might already be available or could easily be found.

Forward selection begins with a model having just a constant term; then a single *x*-variable is added, that which gives the least residual sum of squares; then a second *x*-variable is added to this model in the same way, and this process continues until there is no variable that can be added which significantly reduces the residual sum of squares. This is computationally more efficient than backward elimination. But many combinations of variables remain completely untested: once a particular *x*-variable is in the model, it is never removed and thus an optimal model may not be reached, because there could be a pair (or perhaps a larger set) of *x*-variables which together would have been better even though neither has been selected for the model by itself. Thus a variable already in the model may be retained to the exclusion of other variables that would have been more useful.

Backward elimination starts from the full model containing all variables and removes terms one by one; at each stage the term which makes the least difference in the model sum of squares is removed. Eventually there will be no more terms which can be removed without significantly altering the sum of squares, and the model current at that stage is accepted. This is arguably more valid statistically than forward selection since it starts from the full potential model, but again there is the problem of combinations of variables remaining untested: once a variable has been eliminated it cannot be tried again in a different combination. The process may also suffer from problems resulting from multicollinearity.

(ii) (a) Clearly X_1 enters first, because it gives the largest reduction in the error sum of squares. Once X_1 is there, X_2 is better than X_3 to add to it in the model.

Step 1 (entering X_1) gives a residual SS of 319.12 with 19 df, thus the residual MS is 16.80. So the 1 df reduction in the residual SS by including X_1 is 2069.24 – 319.12 = 1750.12. Thus the usual "extra sum of squares" test statistic is 1750.12/16.80 = 104.2 which on comparing with $F_{1,19}$ is extremely highly significant (the upper 0.1% point is 15.08). So X_1 is retained in the model.

Now adding X_2 gives a further reduction of 319.12 – 188.80 = 130.32, and the residual MS is 188.80/18 = 10.49. The $F_{1,18}$ test statistic is 130.32/10.49 = 12.4 which is significant at 0.5%. So X_2 is also retained in the model.

Adding X_3 to this two-variable model would reduce the residual SS by only 188.80 – 178.83 = 9.97. This is less than the 17 df residual MS of 178.83/17 = 10.52. So we do not add X_3; we stop at X_1 and X_2 and the selected model is $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ + error.

Solution continued on next page

(ii) (b) First, the residual mean square from the full model is 178.83/17 = 10.52 with 17 df.

The smallest change from the full model omits X_3. It increases the residual sum of squares by 188.80 – 178.83 = 9.97. Using the "extra sum of squares" principle, we consider 9.97/10.52 which is clearly not significant on $F_{1,17}$ [**note:** this is of course the same calculation as was done at the end of part (ii)(a)]. This means that the model sum of squares has not been reduced significantly by omitting X_3, so we use this new model (i.e. containing X_1 and X_2 but not X_3) as the basis for the next step.

Omitting X_2 gives the smallest change in the residual sum of squares and therefore also in the model sum of squares (319.12 – 188.80 = 130.32). This is to be compared with the residual mean square from the (X_1, X_2) model which is 188.80/18 = 10.49 with 18 df. So we consider 130.32/10.49 = 12.42, which is significant at the 0.5% level on $F_{1,18}$ [**note:** again this is of course the same as calculation as one already done in part (ii)(a)]. So omitting X_2 would give a significant change, so we do not omit X_2. It follows that we do not omit X_1 either, as that would have given a greater change.

So the selected model is $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ + error.

(ii) (c) Both methods have arrived at the same model, ie that containing X_1 and X_2. The value of R^2 for this model is 91% [(2069.24 – 188.80)/2069.24], giving a further suggestion that it might be a good model. The combination (X_2, X_3) has not been tried in either forward selection or backward elimination, but its residual sum of squares is considerably higher than that for the selected model. X_3 appears to be unimportant whether or not the other variables are present. As far as we can tell from the purely statistical information, the selected model appears to be satisfactory.

(ii) (d) Despite the statistical assurances from part (ii)(c), we cannot be sure that this is a good model. From the statistical point of view, no models with second-order terms (eg squares of variables, or interactions) have been examined; such a model might be better. From the practical point of view, we do not know whether the selected model makes sense – we need to know what the variables X_1, X_2 and X_3 actually are. It would also be desirable to know about the practical conditions under which the data were collected.

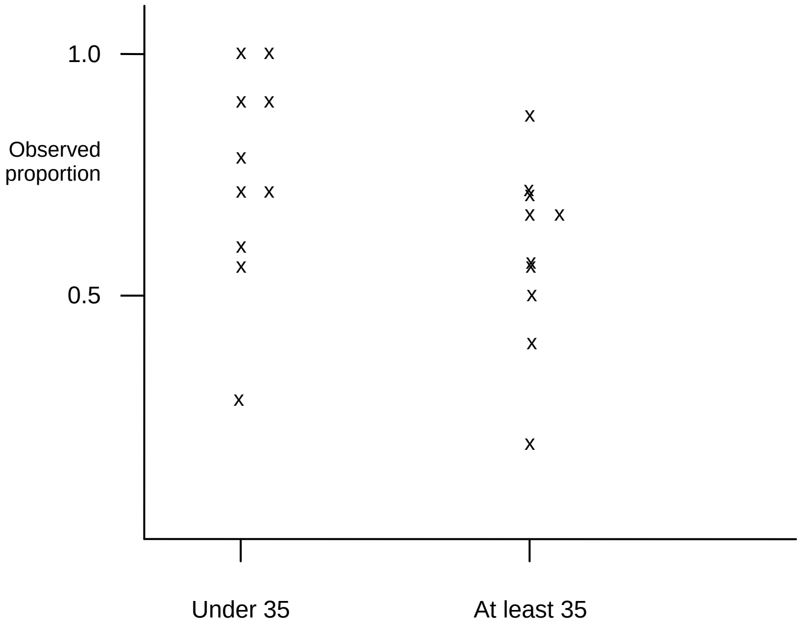
(ii) (e) The usual plots of residuals would be examined – a plot of the residuals themselves to check for constant variance, and plots against fitted values and against predictors to check for (absence of) patterns. The usual regression diagnostics as given by computer programs would also be examined – for example, to check for influential values and for leverage, and to study measures such as Cook's distance.

Graduate Diploma, Module 4, 2009. Question 8

Part (i)

Examine the observed proportions, which are as follows.

Women under 35:	0.90	0.78	0.71	0.60	0.90	1.00	0.56	1.00	0.29	0.71
Women at least 35:	0.86	0.70	0.56	0.50	0.67	0.20	0.57	0.67	0.40	0.71



The observed proportions suggest that the fertilisation rate might be, on the whole, higher for the younger women.

Part (ii)

(a) If, for a particular woman, the fertilisation rate is constant (i.e. the same for each egg) and fertilisation is independent from one egg to another, then a binomial distribution is a valid model. These assumptions may be reasonable from a clinical point of view.

Solution continued on next page

(b) We have $P\left(X_i = x_i \middle| n_i, \pi_i\right) = \binom{n_i}{x_i} \pi_i^{x_i} \left(1-\pi_i\right)^{n_i - x_i}$. Thus the log likelihood is

$$\sum_i \log P\left(X_i = x_i\right)$$

$$= \sum_i \left\{ \log\left(n_i!\right) - \log\left(x_i!\right) - \log\left((n_i - x_i)!\right) + x_i \log \pi_i + \left(n_i - x_i\right) \log\left(1-\pi_i\right) \right\}$$

$$= \sum_i \left\{ \log\left(n_i!\right) - \log\left(x_i!\right) - \log\left((n_i - x_i)!\right) \right\} + \sum_i n_i \log\left(1-\pi_i\right)$$

$$+ \sum_i x_i \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

and the last of these terms is the logit function.

Part (iii)

(a) There are 20 observations of (n_i, x_i) and 2 parameters in the model that has been fitted. So there are 18 degrees of freedom, and the scaled deviance is thus given by deviance/df = 28.26/18 = 1.57. This is a measure of the "fit" of the model. A value near 1 indicates a satisfactory fit – the value here suggests that the fit is acceptable.

(b) V(u) refers to the variance of the binomial distribution. "eggs" is the number of eggs and "u" is the number of fertilised eggs. So the estimate of the binomial proportion is simply u/eggs and the variance is estimated by

$$\text{eggs} \times \frac{\text{u}}{\text{eggs}} \times \left(1 - \frac{\text{u}}{\text{eggs}}\right) = \text{u} \times \left(1 - \frac{\text{u}}{\text{eggs}}\right).$$

g(u) refers to the logit function, log[π/(1 – π)]; inserting the estimate of π gives

$$\log\left(\frac{\text{u}/\text{eggs}}{1-(\text{u}/\text{eggs})}\right) = \log\left(\frac{\text{u}}{\text{eggs}-\text{u}}\right).$$

(c) Recalling that younger women are coded with 0 and older women with 1, the values of the linear predictor are

1.150	for younger women,
1.150 – 0.744 = 0.406	for older women.

Thus the predicted success rates are

$e^{1.150}/(1 + e^{1.150}) = 0.76$	for younger women,
$e^{0.406}/(1 + e^{0.406}) = 0.60$	for older women.

Solution continued on next page

(d) The difference in the scaled deviances is 32.65 – 28.26 = 4.39. The null distribution underlying this quantity (i.e. if including the age effect does not improve the adequacy of the model) is χ^2 with 1 degree of freedom. The upper 5% point of this distribution is 3.84, so the result is significant – the effect of woman's age is statistically significant (at the 5% level), and there is evidence that the age effect should be included in the model.

(e) The change in parametrisation (with age now coded as 1 for younger women and 0 for older women instead of vice versa) makes the new constant plus the age parameter equal to 1.150; and the age parameter is now +0.744; so the constant term in this model is 0.406.

All other results from model-fitting (eg deviances and any *p*-values) will be unaltered. The only difference is the parametrisation.