THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 4

SOLUTIONS FOR SPECIMEN PAPER A

THE QUESTIONS 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.

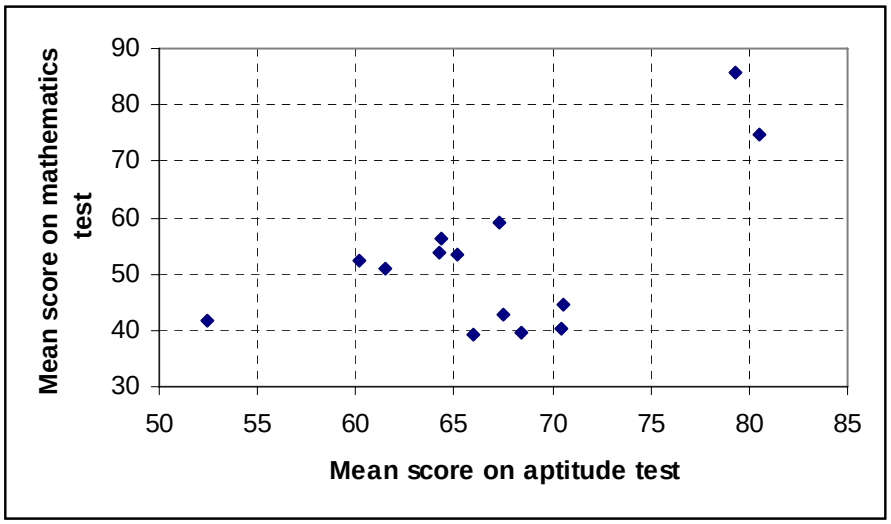
Note. In accordance with the convention used in all the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Graduate Diploma Module 4, Specimen Paper A. Question 1

(i)(a) The least squares estimators of the coefficients of a linear model have minimum variance among all linear unbiased estimators. (This assumes that the residual error terms are uncorrelated random variables with common variance.)

(b) Weighted least squares should be used when the errors are uncorrelated but have different variances, e.g. if the variance is some function of the mean.

(ii)(a)



Note. False "origin" is placed at (50, 30).

It might be useful to add (perhaps in brackets alongside each point) the number of pupils entered by the school.

From the scatter plot, note that the two schools with the smallest number of pupils (these are the schools with aptitude test values near 80) have scores much higher than the others. Otherwise the scatter seems fairly random. However, the two highest points will be very influential in fitting a regression.

(b) Simple linear regression gives a poor fit (R^2 = 38%). The value of "constant" is very poorly determined, though p = 0.019 for the coefficient of "aptitude" seems to suggest some linear relationship. (Fuller output, with information on influence and leverage, would be useful.)

However, the weighted regression, using numbers of pupils as weights, is even less satisfactory, giving an even lower R^2 and no evidence of a linear relationship (p = 0.193).

The mean maths score has variance σ_i^2/n_i where σ_i^2 is the within-school variance. The weighting assumes that all the σ_i^2 are similar, because then n_i is a suitable weight. But the σ_i^2 are not likely to be (approximately) equal, and until we have all the individual marks we cannot obtain the alternative weighting factors n_i / σ_i^2. (For example, the schools with small n_i might have selected pupils, leading to smaller σ_i^2 than the others.)

Neither regression is adequate, although the unweighted one might be a fair reflection of what is seen from the graph. Without more "diagnostic" information, we cannot go any further.

Graduate Diploma Module 4, Specimen Paper A. Question 2

(i) The events are rare, if the system is run by experienced people, and they may be assumed to be random; if so, the Poisson is the appropriate distribution. The log link function is the natural one for a Poisson distribution.

(ii) There are 12 observations, and 3 estimated parameters. The scale parameter is 1. Hence the deviance has 9 d.f. We have

$$\frac{\text{deviance}}{\text{d.f.}} = \frac{11.96}{9} = 1.33,$$

quite near to 1. On the basis of deviance, the fit looks reasonable. But there are other criteria to consider.

(iii)
1. The plot of residuals against predicted values is not random, but shows a somewhat parabolic trend.
2. The plot against SPEC1 is a divergent linear trend.
3. The plot against SPEC2 shows a fan shape.

These all indicate poor fit of the model. Linear functions of SPEC1 and SPEC2 may not be appropriate.

(iv) The histogram is very skew. The "Normal plot" is curved. The Kolmogorov-Smirnov test has $p \approx 0.08$, which is not good. The evidence points to non-Normal residuals. This again suggests a poor fit of the model.

(v) The motivation for this may have been to compress the SPEC1 scale of measurement, and was probably reasonable in view of the residual plots from the original.

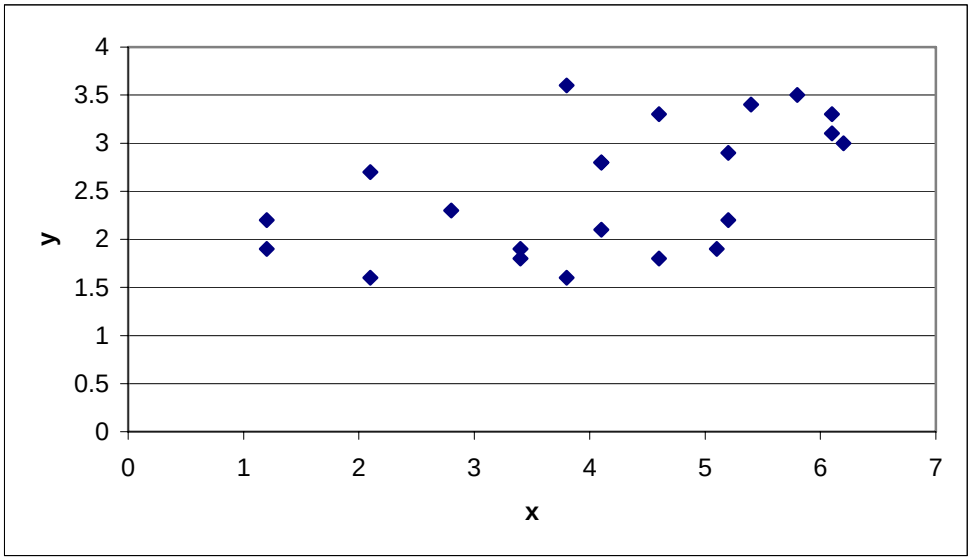
(vi) With 9 d.f., deviance/d.f. = 0.48, much less than 1, so this model appears to be a much better fit.

(vii)
1. The plot of residuals against fitted values looks somewhat better, although there is still a hint of curvature.
2. The plot against SPEC1 is now satisfactory.
3. The plot against SPEC2 is similar to the previous one. However, the histogram and the Normal plot show a better degree of Normality than before.

(viii) There may be no reason to specify what transformation (if any) is appropriate for SPEC1 and SPEC2; therefore it will be worth trying something like √ for SPEC2, though it may be suitable to retain log for SPEC1. The residual plots give little further guide to this. Deviance is only a guide to the fit of a model, and the residual plots are a useful addition. No obvious, simple-to-interpret, model seems to exist.

Graduate Diploma Module 4, Specimen Paper A. Question 3

(i)



There is a considerable amount of scatter but some indication of a weak positive association between y and x. The variance of the y variable looks as if it could be assumed constant – there is no apparent pattern (such as a dependence on x).

(ii) Any association that exists between y and x is not obviously curved, and does appear to have a linear component. So a linear regression model seems reasonable. Because there are repeat observations on y at some of the x-values, a "pure error" term can be extracted from the residual as the sum of squares between these repeats (see below). The remainder of the residual ("lack of fit") then represents departure from linearity, which can be tested against the "pure error". This should give a better test of the linear regression model.

The model is

$$Y_{ij} = a + bx_i + \varepsilon_{ij} \qquad i = 1, 2, \ldots, 13$$

$j = 1$ or 2 or 3, depending on the value of i.

where

x_i is a value of x,

Y_{ij} represent the (single or repeat) observations taken at $x = x_i$,

$\{\varepsilon_{ij}\}$ are independent normal N(0, σ^2) random variation terms with constant variance.

Solution continued on next page

(iii) (a) At x = 1.2, we have y = 2.2 and 1.9, with total 4.1. So the pure error SS here is $2.2^2 + 1.9^2 - \frac{4.1^2}{2} = 0.045$. This has 1 degree of freedom.

(b) At x = 4.1, we have y = 2.8, 2.8 and 2.1, with total 7.7. So the pure error SS here is $2.8^2 + 2.8^2 + 2.1^2 - \frac{7.7^2}{3} = 0.327$. This has 2 df.

(c) At each x value where there are repeats, a similar calculation is carried out. The sums of squares are added to obtain the total pure error SS. The numbers of degrees of freedom would also be added to obtain the total df for "pure error", which here will be 9 (this is needed below, in part (iv)).

(iv) If the "pure error" SS is 4.3717, the "lack of fit" SS must be 6.6554 – 4.3717 = 2.2837, and this will have 20 – 9 = 11 df.

Hence the analysis of variance is

Source of variation	df	Sum of squares	Mean square	F value
Regression	1	2.6723	2.6723	2.6723/0.4857 = 5.50
Lack of fit	11	2.2837	0.2076	0.2076/0.4857 = 0.43
Pure error	9	4.3717	0.4857	= $\hat{\sigma}^2$
Total	21	9.3277		

The F value for regression (note that this is now a comparison with the *pure error* term) is referred to the $F_{1,9}$ distribution. This is significant at the 5% level (critical point is 5.12), so there is some evidence in favour of the regression model.

There is no evidence of lack of fit. (It could be argued that the lack of fit and pure error SSs should therefore be recombined to give the residual as before, with 20 df.)

We note that R^2 = 2.6723/9.3277 = 28.6%, which is low; despite the absence of evidence for lack of fit, only about 29% of the variation in the data is explained by the linear regression model. This is because the underlying variability (estimated by the pure error mean square) is high.

(v) Residuals after fitting the proposed model can be examined, and any patterns in them noted. Departures from the model can be detected in this way, such as a need for an additional term, or systematic non-constant variance, etc.

Graduate Diploma Module 4, Specimen Paper A. Question 4

(i) 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. As shown in part (ii), a partial F test is used to check this. 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.

Disadvantages are that the method works in an "automatic" way which does not use knowledge about what the variables actually are; and that once a variable has been eliminated it cannot be tried again in a different combination (as is done by the "all possible regressions" method).

It may be preferred to forward selection since it does necessarily include all the variables at the beginning of the process, whereas forward selection may not test some of the variables (even some that may in fact be important) at all.

Another advantage is that although it begins with the "full" model, it does not require so much computing as the "all possible regressions" method.

Multicollinearity remains a problem with backward elimination.

(ii) [Note. This is a rather small data set for this purpose.]

First, the residual sum of squares from the full model is 2715.76 – 2667.90 = 47.86 with 8 df, so we initially take 47.86/8 = 5.9825 as the residual mean square.

The smallest change from the full model omits X_3. It reduces the model SS by 0.11. Using the "extra sum of squares" principle, we consider 0.11/5.9825 which is approximately 0.02 and clearly not significant on $F_{1,8}$. This means that the model sum of squares has not been reduced significantly, so we use this new model (i.e. containing X_1, X_2 and X_4) as the basis for the next step.

Omitting X_4 gives the smallest change in the model sum of squares (2667.79 – 2657.90 = 9.89). This is to be compared with the residual from the (X_1, X_2, X_4) model which is 2715.76 – 2667.79 = 47.97 with 9 df. So we consider 9.89/(47.97/9) = 1.86, not significant on $F_{1,9}$. So we now consider the (X_1, X_2) model.

The smallest change is by removal of X_2, the change being 2657.90 – 1809.40 = 848.5. This should be compared with the residual from the (X_1, X_2) model, which is 2715.76 – 2657.90 = 57.86 with 10 df. So we consider 848.5/(5.786) = 146.6, which is extremely highly significant on $F_{1,10}$. Thus we do *not* remove X_2, and the final model is (X_1, X_2).

Solution continued on next page

(iii) Any existing knowledge of relations between Y and the Xs is valuable (especially when given only a small data set, as here). We should not operate merely from the sums of squares alone.

Note that the first step in the above method showed very little to choose between three of the 3-variable models. Similarly for the final model the sums of squares show little to choose between (X_1, X_2) and (X_1, X_4); indeed, (X_2, X_3) also looks worthy of consideration even though X_3 had been eliminated in the first step. Note also that forward selection would have started with X_4 – but this was eliminated in the backward selection!

There are likely to be correlations among the Xs which could indicate that some pairs are giving almost the same information – possibly X_2 and X_4 in this example. A correlation matrix or scatter diagrams will often help in deciding how to proceed.

There is also the point that some variables may be easier and quicker to measure, or known to be more reliable.

(iv) (a) The statement is rather over-emphatic but contains good sense. For a large set of data, results should not be "wildly" wrong; but in all cases the above discussion (part (iii)) is relevant. It is good practice to encourage an approach that is not purely automatic/arithmetic but also practical, especially when a manuscript covers just one stage in a programme of work.

(b) Various regression diagnostics are available in computer packages. Study of the residuals can reveal possible outliers which are unduly influencing results as well as checking for the Normality of residuals (by use of a Normal probability plot) that is assumed in *F* tests. For particular types of work (eg time series), particular methods are commonly used; likewise, Durbin-Watson tests are commonly used in econometrics.

Graduate Diploma Module 4, Specimen Paper A. Question 5

(i) Driver totals are: *A* 173, *B* 151, *C* 201, *D* 163.

"Correction factor" is $\frac{688^2}{16} = 29584$. Therefore total SS = 30042 – 29584 = 458.

SS for drivers = $\frac{173^2}{4} + \frac{151^2}{4} + \frac{201^2}{4} + \frac{163^2}{4} - 29584 = 29925 - 29584 = 341$.

SS for cars = $\frac{181^2}{4} + \frac{171^2}{4} + \frac{161^2}{4} + \frac{175^2}{4} - 29584 = 29637 - 29584 = 53$.

SS for roads = $\frac{182^2}{4} + \frac{174^2}{4} + \frac{164^2}{4} + \frac{168^2}{4} - 29584 = 29630 - 29584 = 46$.

Hence:

SOURCE	DF	SS	MS	*F* value
Cars	3	53	17.67	5.89 significant
Roads	3	46	15.33	5.11 significant
Drivers	3	341	113.67	37.89 very highly sig
Residual	6	18	3.00	$= \hat{\sigma}^2$
TOTAL	15	458		

[All *F* values are compared with $F_{3,6}$; upper 5% point is 4.76, upper 0.1% point is 23.7.]

There are differences between cars and between roads, both significant at the 5% level; these might not look very large differences, but the residual error variability, with which they are compared, is quite small. The difference between drivers is much stronger – significant at the 0.1% level – with driver *C* having a relatively large value.

(ii) Combinations of all cars with all roads and all drivers would require 4×4×4 = 64 runs. The Latin square scheme, in 16 runs, allows orthogonal comparisons of the three factors, on the assumption that there are no interactions. A 4×4 square has only 6 degrees of freedom for residual, and often that would not be enough to give a reliable estimate of σ^2; here, however, the estimate is quite small, so a useful analysis has resulted. Using two squares would give ample degrees of freedom for *F* and *t* tests.

(iii) There are four "standard" 4×4 squares (letters in alphabetical order in first row and in first column), one of which must be 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* are then allocated at random to the "treatments" (drivers). This gives a random choice from all possible 4×4 squares.

Solution continued on next page

(iv) Note that *t* tests would show little difference among *A*, *B*, *D* but a significantly greater amount of wear when *C* is driving.

Contrasts:

	A	*B*	*C*	*D*	Value	Divisor	SS	*F* value
TOTAL	173	151	201	163				
Times of day	–1	1	–1	1	–60	16	225	75.00
Weekday/weekend	–1	–1	1	1	40	16	100	33.33
Interaction	1	–1	–1	1	–16	16	16	5.33

The *F* values are all compared with $F_{1,6}$; upper 5% point is 5.99, upper 1% point is 13.74, upper 0.1% point is 35.51. Thus the result for time of day is very highly significant, that for weekday/weekend is highly significant, and that for interaction is significant.

Morning times (*A*, *C*) give a great deal heavier wear; so do weekdays. But since *C* is different from the others, and *C* drove on weekday mornings, this may explain all of these results; we cannot give any firm conclusions.

Graduate Diploma Module 4, Specimen Paper A. Question 6

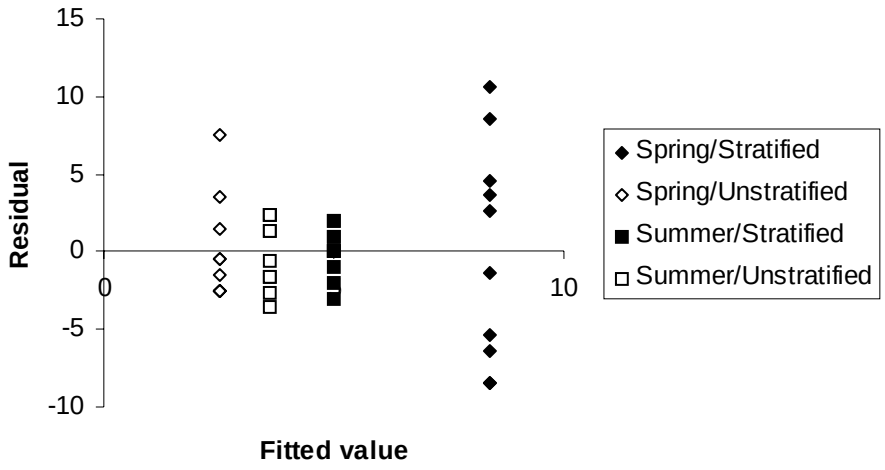
[solution continues on next page]

(i) In a one-way analysis of variance, the residual for any plot is the difference between the observed value and the fitted value, which is simply the mean for that treatment. For example, the mean for Spring/Stratified is 8.4, so that the residual for observation '12' is 12 – 8.4 = 3.6. The sum of the residuals for this treatment, and for each of the other treatments, will of course be 0.

Before carrying out further analysis, note that each data item should actually have an underlying binomial distribution with $n = 20$. There are a number of extreme values, near to 0 or 20. The ranges of the data for the four treatments are Spring/Stratified 0 to 19, Spring/Unstratified 0 to 10, Summer/Stratified 2 to 7, Summer/Unstratified 0 to 6. All this suggests that the required assumptions of underlying Normality and equal variances for the four treatments seem unlikely to be the met. An angular transformation may be necessary to stabilise variance.

The residuals can be plotted against the fitted values. The plot is shown below. The pattern of the scatter should be the same for each treatment. It does not appear to be so. For example, spring storage gives more variable results than summer storage.

[**Note.** There are coincident points for each fitted value. The full list of residuals is given in the question.]



A Normal probability plot would be possible if a computer is available. (It gives some evidence of non-Normality due to the extreme values; these are also noticeable in the plot above, though not in a way that suggests lack of symmetry.)

[Candidates might mention Bartlett's test for variance homogeneity; but it should be emphasised that it is sensitive to non-Normality and so in this case is unlikely to be useful.]

Since the layout is effectively "completely randomised", there are no classifications other than treatments that can be studied.

(ii) The assumptions are that the (true) residuals (experimental errors) are i.i.d. (independent identically distributed) N(0, σ^2) and that there is no systematic variation except treatments.

The required discussion is included in the solution to part (i) above. An angular transformation is suggested, but even then we may not have good results because of the different patterns of scatter within the treatments. The usual F and t tests might well be unreliable. Individual comparisons between treatments could be made, not using the overall estimate of variance from the ANOVA; or a non-parametric comparison could be made.

Graduate Diploma Module 4, Specimen Paper A. Question 7

(i) The grand total is 127.03 the "correction factor" is $127.03^2/60 = 268.9437$.

So the total sum of squares = $301.4107 - \frac{127.03^2}{60} = 32.4670$, with 59 df.

SS for blocks = $\frac{36.09^2}{20} + \frac{43.27^2}{20} + \frac{47.67^2}{20} - \frac{127.03^2}{60} = 272.3605 - 268.9437$

= 3.4168, with 2 df.

SS for seed rate = $\frac{10.92^2}{12} + ... + \frac{31.64^2}{12} - 268.9437 = 25.6476$, with 4 df.

SS for row width = $\frac{31.48^2}{15} + ... + \frac{29.01^2}{15} - 268.9437 = 0.9166$, with 3 df.

Interaction SS = $\frac{1.87^2}{3} + \frac{5.40^2}{3} + ... + \frac{7.05^2}{3} - 268.9437 - 25.6476 - 0.9166$

= 0.9976, with $4 \times 3 = 12$ df.

The residual SS and df follow by subtraction.

Hence:

SOURCE	DF	SS	MS	*F* value
Blocks	2	3.4168	1.7084	
Seed rate	4	25.6476	6.4119	163.6
Row width	3	0.9166	0.3055	7.8
Interaction	12	0.9976	0.0831	2.1
Residual	38	1.4884	0.0392	$= \hat{\sigma}^2$
TOTAL	59	32.4670		

The *F* value of 163.6 is referred to $F_{4,38}$; this is well beyond the upper 0.1% point (about 5.8), so there is extremely strong evidence of an effect of seed rate.

The *F* value of 7.8 is referred to $F_{3,38}$; this is beyond the upper 0.1% point (about 6.7), so there is very strong evidence of an effect of row width.

The *F* value of 2.1 is referred to $F_{12,38}$; this is (just) significant at the 5% level, so there is some evidence of an interaction.

Overall, though the effects of seed rate and row width appear very highly significant, the results should be explained in terms of the interaction.

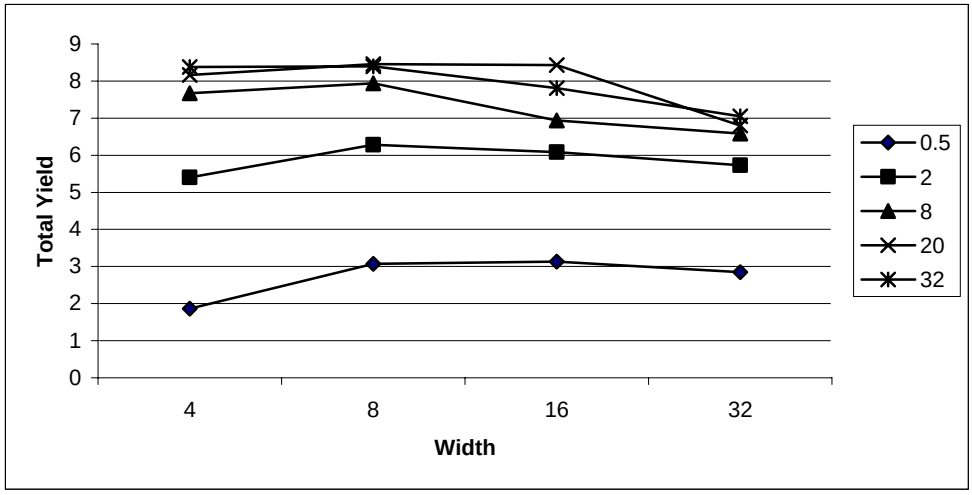
Solution continued on next page

(ii) The partitioning for row width is carried out as follows.

Row width	4	8	16	32	r = 15 (for each total)			
Total	31.48	34.15	32.39	29.01	Value	Divisor	SS	F value
Linear	–3	–1	1	3	–9.17	15×20	0.2803	7.2
Quadratic	1	–1	–1	1	–6.05	15×4	0.6100	15.6
Cubic	–1	3	–3	1	2.81	15×20	0.0263	0.7
							0.9166	

Each partitioned term has 1 df and F tests (comparing with the residual mean square as before) have 1 and 38 df, so there is evidence for a linear component of the effect and very strong evidence for a quadratic component.

(iii)



(iv) Looked at overall, the yield is greatest for row width 8; the yield from row widths 4 and 16 are close together at a somewhat lesser value, and the yield from row width 32 is the least. The rise and fall with respect to row width leads to the quadratic component of this main effect.

In terms of overall effect of seed rate, the yield from rate 0.5 is very much less than that from rate 2 which is itself substantially less than that from the others.

However, the interaction has to be taken into account. The diagram shows that there are somewhat different patterns of yields over the row widths at the different seed rates.

A seed rate of 8lb/acre seems adequate and is likely to be economical, although any future work needs to clarify the row width appropriate for this seed rate to give maximum yield.

Graduate Diploma Module 4, Specimen Paper A. Question 8

(a) (i) All combinations of all levels of the factors are used as the set of "treatments". Here "factors" are Varieties, Soil types, Moisture level. If interactions among factors may be present, examining one factor at a time will not give valid results for inferring what happens when they are used together. Even when factors do not interact, they have been examined over a wide range of conditions and the results should have more general validity.

(ii) $$\text{VSM} = \frac{1}{4r}(v-1)(s-1)(m-1) = \frac{vsm - vs - vm - sm + v + s + m - (1)}{4r}$$

$$= \frac{(182 - 153 - 131 - 113 + 122 + 98 + 96 - 50)}{(4 \times 4)} = \frac{51}{16} = 3.1875.$$

[There are 4 comparisons of (+, –) to be arranged.]

(iii) $\sum y = 945$, correction term is $\frac{945^2}{32} = 27907.03125$.

Total SS is therefore 3605.9688.

Blocks SS = $\frac{1}{8}\left(218^2 + 256^2 + 212^2 + 259^2\right) - \frac{945^2}{32} = 228.5938$.

Each factorial term has SS = $2r \times (\text{effect estimate})^2 = 8(\text{effect}^2)$.

Hence:

SOURCE	DF	SS	MS	*F* value
Greenhouses	3	228.5938	76.198	2.59 not significant
V	1	1667.5313	1667.531	56.70 very highly sig
S	1	675.2813	675.281	22.96 very highly sig
M	1	306.2813	306.281	10.41 highly sig
VS	1	9.0313	9.031	0.31 not significant
VM	1	16.5313	16.531	0.56 not significant
SM	1	3.7813	3.781	0.13 not significant
VSM	1	81.2813	81.281	2.76 not significant
Residual	21	617.6559	29.4122	$= \hat{\sigma}^2$
TOTAL	31	3605.9688		

[NOTE: SS values are given to greater accuracy here than is possible from the information given on the paper.]

Testing each 1 d.f. MS against the residual, we find highly significant main effects but no significant interactions. The higher yields were obtained when V_2 was used, grown in S_2, at high moisture level M_2.

Solution continued on next page

(b) (i) The block size is now smaller than the number of treatments used. The blocks (greenhouses) can each only contain half of the full set of treatments. Thus we cannot obtain information on all possible treatment effects from any one block, and any consistent differences that might exist between the blocks would confuse (*confound*) the treatment effect comparisons. However, eight blocks are available and, if there is a high-order interaction which is believed to be unimportant, it can be arranged that it has the same pattern of $\pm$ signs as a comparison between two blocks. In these circumstances, blocking can therefore still be used to take out greenhouse differences without necessarily losing important information from the experiment.

(ii) We might sensibly choose to confound VSM, believing (or at least hoping) that, being the three-factor interaction, it is the least likely to be important. The treatments *v*, *s*, *m*, *vsm* would then be placed in random order in one greenhouse, and (1), *vs*, *vm*, *sm* in another greenhouse (perhaps the adjacent one). The comparison between these two greenhouses would be measured as part of the blocks SS, and the treatment effect VSM would thus be confounded with blocks. The same procedure could be used in each of four pairs of greenhouses.