THE ROYAL STATISTICAL SOCIETY

2001 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

APPLIED STATISTICS

PAPER I

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.

Graduate Diploma, Applied Statistics, Paper I, 2001. Question 1.

(a)

(i) If **X** is the "design matrix", **X'X** will be singular because the model contains redundant parameters.

(ii) One method is to omit one of the *a* variables and one of the *b* variables.

(iii) The table of cell means is:

	a_1	a_2	a_3
b_1	$\mu + \gamma_1 + \beta_1$	$\mu + \gamma_2 + \beta_1$	$\mu + \beta_1$
b_2	$\mu + \gamma_1 + \beta_2$	$\mu + \gamma_2 + \beta_2$	$\mu + \beta_2$
b_3	$\mu + \gamma_1$	$\mu + \gamma_2$	μ

So μ is estimated by the mean at a_3b_3. Then γ_2 is the difference between this and the mean at a_2b_3 or between a_3b_2 and a_2b_2 or between a_3b_1 and a_2b_1; obviously the mean of these three differences will be used, i.e. γ_2 is the difference in the (marginal) means of *A* at levels 2 and 3. The remaining parameters can be estimated in similar ways.

(iv) Additional dummy variables c_{ij} are required, where $c_{ij} = a_ib_j$. We only require c_{11}, c_{12}, c_{21}, c_{22}. The model therefore becomes

$$y = \mu + \gamma_1 a_1 + \gamma_2 a_2 + \beta_1 b_1 + \beta_2 b_2 + \kappa_{11} c_{11} + \kappa_{12} c_{12} + \kappa_{21} c_{21} + \kappa_{22} c_{22} + e\,.$$

Assuming there are replicates in each cell, the reduction in sum of squares through fitting the extended model is a measure of interaction which can be tested for significance.

(b)

(i) The variables should be re-scaled, e.g. standardised (or possibly transformed). We require no single variable to dominate simply because of its scale of measurement.

(ii) Either omit one or more variables that are very highly correlated with other variables in the model – so avoid singularity problems;

or try new combinations of the original variables, such as principal components, which may better reflect the important aspects of the data;

or use ridge regression;

or use a generalised inverse of **X'X** to get over singularity problems;

or identify the likely causes of the problem by calculating variance inflation factors and an eigenanalysis of **X'X**.

(iii) Use weighted least squares, with weights inversely proportional to the variances of the residuals. Some initial guesswork and an iterative procedure will be required.

(iv) This problem is generally due to multicollinearity, so the ideas in (ii) could be used. If the significance level is rather weak, the model may not in any case be very useful.

Graduate Diploma, Applied Statistics, Paper I, 2001. Question 2

(i) BACKWARD SELECTION. Begin with AGE+HEIG+WEIG+CHES, and delete one variable, in turn:

Term deleted	F ratio		
AGE	$\frac{(0.014175-0.0069226)}{\frac{0.0069226}{5}}=5.24$	(1, 5 df)	
HEIG	$\frac{(0.10249-0.0069226)}{\frac{0.0069226}{5}}=69.02$		
WEIG	$\frac{(0.011133-0.0069226)}{\frac{0.0069226}{5}}=3.04$		
CHES	$\frac{(0.0071493-0.0069226)}{\frac{0.0069226}{5}}=0.164$		delete

Repeat, beginning from AGE+HEIG+WEIG:

AGE	$\frac{(0.01516-0.0071493)}{\frac{0.0071493}{6}}=6.72$	(1, 6 df)	
HEIG	$\frac{(0.12079-0.0071493)}{\frac{0.0071493}{6}}=95.37$		
WEIG	$\frac{(0.012758-0.0071493)}{\frac{0.0071493}{6}}=4.71$		delete

Using AGE+HEIG:

AGE	$\frac{(0.020102-0.012758)}{\frac{0.012758}{7}}=4.03$	(1, 7 df)	delete
HEIG	$\frac{(0.20901-0.012758)}{\frac{0.012758}{7}}=107.68$		

Finally, remove HEIG:

HEIG	$\frac{(0.21296-0.020102)}{\frac{0.020102}{8}}=76.75$	(1, 8 df)

Retain HEIG. It is the only term which is significant. None of those deleted at any stage reached the 5% level.

(ii) There is one observation (8) with extremely large influence relative to the others, and also high leverage. Another (5) also has moderate leverage.

1. The data should be checked for accuracy, and if the subjects are available should be re-measured.

2. Are any of the subjects "unusual", giving outlying results (even if correct ones)?

3. Repeat the analysis leaving out the point(s) which have been highlighted in the diagnostic tests. Examine the effect of doing this, and interpret these results in the context of the problem.

4. Other influence statistics may be considered, reflecting different aspects of the data.

Graduate Diploma, Applied Statistics, Paper I, 2001. Question 3

(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, Applied Statistics, Paper I, 2001. Question 4

(i) There is strong evidence of a linear relationship between density and pressure, with possible slight doubt about variance homogeneity.

(ii) Either R^2 = 98.2% says that all but 1.8% of the variation in density can be explained by its linear relation with pressure – this is not an unbiased estimate;

Or Adj R^2 = 98.1% says that all but 1.9% of the variation in density can be so explained – this is an unbiased estimate, and whereas R^2 must increase when an extra term is added to a model, adjusted R^2 need not because it contains a correction for degrees of freedom.

[Note: Mallows' C_p is often used instead of adjusted R^2.]

(iii) No. The value of the gradient depends on the scales of measurement used for the variables. However, even if the scale for pressure were, say, thousands psi, a change of one unit of pressure would still have very small effects on the density because the value of the constant (2.37) dominates this. But the standard error of the parameter estimate is very small also, and the t value very highly significant, so that there is a good linear relation.

(iv) Assumptions:

1. Homogeneity of variance: some doubt, but the 8000 pressure level is the main reason for this, others being satisfactory.
2. Normality: no evidence against this, either from the histogram or the Normal plot or the statistical test.
3. Independence: no information on which to judge.
4. Linearity: plot of residuals versus fitted values show no curvature, so the model is probably satisfactory and does not need extra terms added.

   A homogeneity of variance test may not be worth doing as it lacks sensitivity.

(v) (a) Replicates of the observations at the five levels of pressure allow a "pure error" term to be found, against which "lack of fit" can be tested.

(b) One method is to fit a model treating the predictor variable as a factor, which will yield independent estimates of σ^2 and the non-linearity deviations. Alternatively, find the variance within each factor level, and pool these estimates; remove this from the general 'deviations' d.f.

Graduate Diploma, Applied Statistics, Paper I, 2001. Question 5

(i) There is a weak linear relation between *PLANTP* and *INORGP*, with an outlying value; *PLANTP* and *ORGP* seem to have no relation when this outlier has been removed. *INORGP* and *ORGP* have a weak linear relation with considerable scatter, so inclusion of both these in a model could lead to multicollinearity; care in interpretation will be needed.

(ii) From the table at the top of Appendix Page 10, the single variable *INORGP* is the best predictor, alone; adding *ORGP* slightly reduces the amount of variation explained. (*ORGP* alone is very poor.) A good model has most of the variation in *PLANTP* explained, thus a high percentage R^2 or adj R^2.

C_p gives a measure of the overall goodness of fit of a model, based on the number of parameters fitted. A model with C_p close to or below the number of parameters (including the constant) in the model is satisfactory. In this example, $C_p = 1$ for *INORGP* alone is good.

(iii) (a) $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i; \quad \{\varepsilon_i\}$ *i.i.d.* $\mathrm{N}(0, \sigma^2)$ for $i = 1, 2, ..., n.$

β_0 is the intercept parameter ("constant"), β_1 the coefficient of the predictor variable *INORGP* and ε_i the residual.

(b) $\mathrm{H}_0 : \beta_1 = 0,$ given $\beta_0 \neq 0; \quad \mathrm{H}_1 : \beta_1 \neq 0,$ given $\beta_0 \neq 0$.

The *p*-value is ≈ 0.001, very highly significant, so there is strong evidence against H_0 and in favour of using x_1 in the model.

(iv) Comparing Appendix Pages 10 and 11, removing 17, the observation with very high *PLANTP* value, has reduced the gradient, increased the intercept, considerably reduced the standard errors of the intercept and gradient estimates, reduced $\hat{\sigma}^2$ considerably, increased by a small extent the amount of variation explained by the relation, and highlighted the observation (10) with the lowest *PLANTP* value (but not otherwise an 'outlier').

Observation 17 is thus highly influential, as could be seen from the original scatter plots. (If the data were still available for checking it would be worth doing so.)

Graduate Diploma, Applied Statistics, Paper I, 2001. Question 6

(a) $E[X_t] = E[Z_t] + 0.7E[Z_{t-1}] - 0.2E[Z_{t-2}] = 0$ since $E[Z_t] \equiv 0 \; \forall t$.

$\text{Var}[X_t] = \text{Var}[Z_t] + (0.7)^2 \text{Var}[Z_{t-1}] + (0.2)^2 \text{Var}[Z_{t-2}]$ since $\{Z_t\}$ independent

$= \sigma_z^2 (1 + 0.49 + 0.04) = 1.53\sigma_z^2 \equiv \gamma_X(0)$.

$$\begin{aligned}\text{Cov}[X_t, X_{t-1}] &= \text{Cov}\left[(Z_t + 0.7Z_{t-1} - 0.2Z_{t-2}), (Z_{t-1} + 0.7Z_{t-2} - 0.2Z_{t-3})\right] \\ &= 0.7\,\text{Var}[Z_{t-1}] - 0.14\,\text{Var}[Z_{t-2}] \qquad \text{by independence of } \{Z_t\} \\ &= 0.56\sigma_z^2 \equiv \gamma_X(1).\end{aligned}$$

$\text{Cov}[X_t, X_{t-2}] = -0.2\text{Var}[Z_{t-2}] = -0.2\sigma_z^2 \equiv \gamma_X(2)$

and clearly $\text{Cov}[X_t, X_{t-k}] = 0$ for all $k \geq 3$.

Hence $\gamma_X(k) = \begin{cases} 1.53\sigma_z^2 & (k=0) \\ 0.56\sigma_z^2 & (k=1) \\ -0.20\sigma_z^2 & (k=2) \\ 0 & (k>2) \end{cases}$

and the process is second order stationary because the mean is stationary and the autocovariance function depends only on the lag separation between the elements in the process.

(b) (i) There should be no obvious structure, with approximately 95% of coefficients in the range $\pm\frac{2\sigma}{\sqrt{n}}$.

(ii) Coefficients will alternate in sign and decrease in magnitude: γ_1 large negative, γ_2 smaller positive, and so on.

(iii) Coefficients will decrease very slowly.

(iv) Coefficients will be dominated by oscillations at the same frequency as that of the seasonal fluctuations.

Graduate Diploma, Applied Statistics, Paper I, 2001. Question 7

(a) (i) $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

where μ = grand mean,

α_i = deviation from grand mean due to main effect of A (i = 1, 2, 3),

β_j = main effect of B in the same way (j = 1, 2),

$(\alpha\beta)_{ij}$ is an interaction term specific to levels i of A and j of B,

ε_{ijk} are i.i.d. N(0, σ^2) random variables,

y_{ijk} = response on kth replicate of (ij)th combination of A and B.

(ii)

TOTALS

	A1	*A2*	*A3*		
B1	38	50	58	: 146	$\Sigma y_{ijk}^2 = 2142$
B2	36	32	26	: 94	$N = 30$
	74	82	84	240 ≡ G	

Correction term $G^2/N = 1920$. SS Total = 2142 – 1920 = 222.

$$SSA = \frac{1}{10}\left(74^2 + 82^2 + 84^2\right) - 1920 = 5.6 .$$

$$SSB = \frac{1}{15}\left(146^2 + 94^2\right) - 1920 = 90.133 .$$

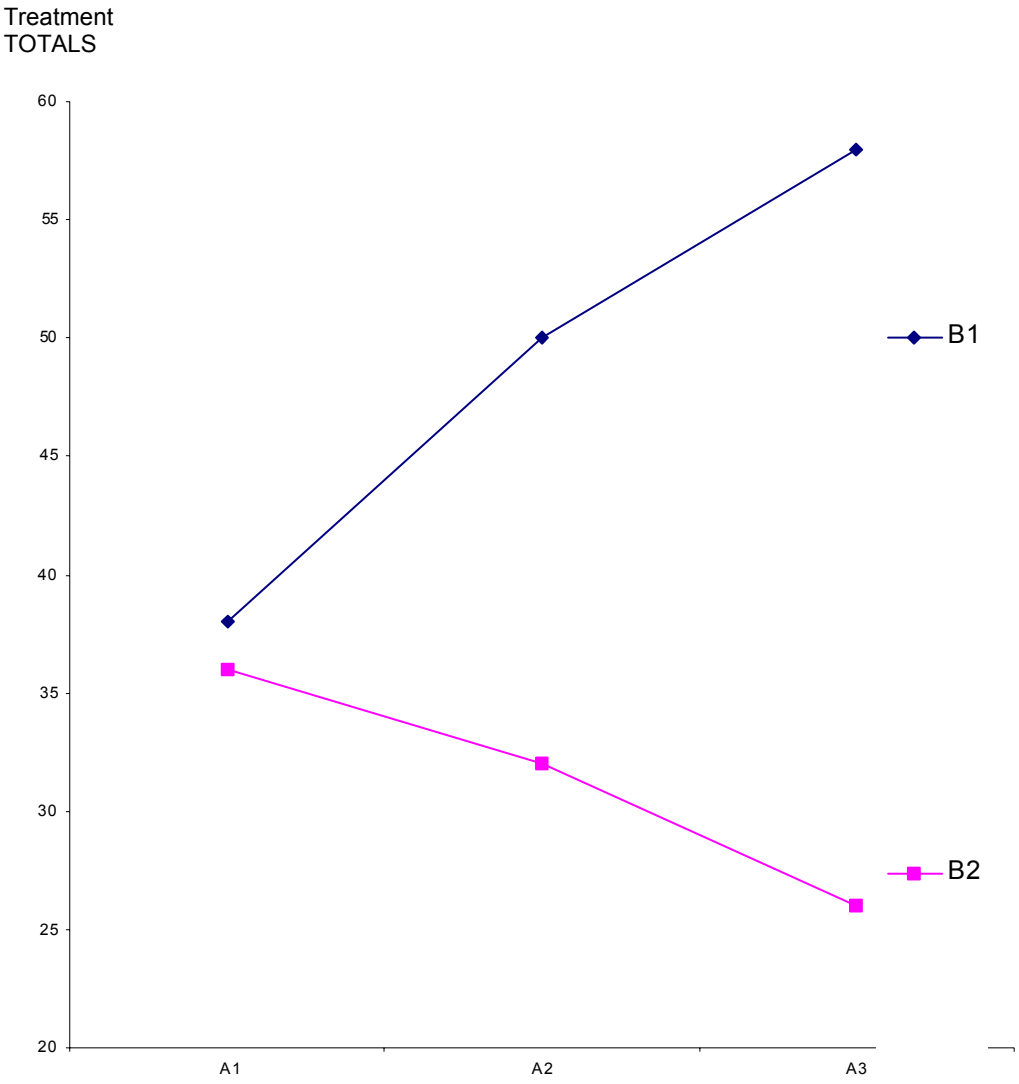
$$SS\left(A + B + AB\right) = \frac{1}{5}\left(38^2 + 50^2 + 58^2 + 36^2 + 32^2 + 26^2\right) - 1920 = 140.8 .$$

Analysis of Variance

ITEM	DF	SS	MS	
A	2	5.600	2.800	
B	1	90.133	90.133	
AB	2	45.067	22.534	$F_{2,24} = 6.66$
	5	140.800		
Residual	24	81.200	3.383	
TOTAL	29	222.000		

Comparing 6.66 with $F_{2,24}$ gives a highly significant result.

(iii)



Because of the interaction, the response to *A* is completely different at the two levels of *B*. (Main effects have no useful meaning.) At *B1*, there is a considerable increase in response from *A1* to *A2*, and again, slightly smaller, from *A2* to *A3*. At *B2* there is a steady decrease from *A1* to *A2* to *A3*.

(b) A 'fixed' factor is one whose experimental levels are determined in advance, and inference is to be made concerning these levels only. A 'random' effect is one where the actual 'levels' (e.g. the laboratories taking part in a study) are a random sample from a wider population to which inferences will be extended.

The terms for *B* and interaction are now taken as $\mathrm{N}\left(0,\sigma_B^2\right)$ and $\mathrm{N}\left(0,\sigma_{AB}^2\right)$, while the residual estimates σ^2, the variance of an individual observation. For the given example, the mean squares estimate the following:

A	$\sigma^2+5\sigma_{AB}^2$	(on the usual NH that fixed effect = 0)
B	$\sigma^2+5\sigma_{AB}^2+15\sigma_B^2$	
AB	$\sigma^2+5\sigma_{AB}^2$	
Residual	σ^2	

On an alternative hypothesis of the usual form, the usual additional term in $\sum\alpha_i^2$ appears in *A*. Estimates of σ^2, σ_{AB}^2 and σ_B^2 are found from the last three rows, and a test of "all $\alpha_i = 0$" is made from the top row against *AB*: clearly here the *A* term is very small, and cannot sensibly be used.

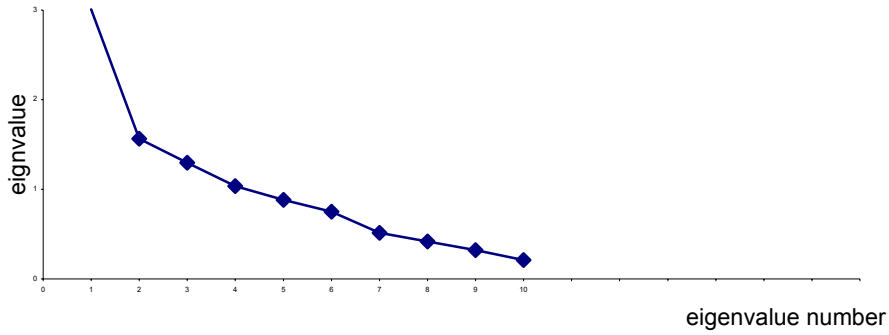
Graduate Diploma, Applied Statistics, Paper I, 2001. Question 8

(a) PCA produces uncorrelated components which are weighted linear combinations of p measurements made on each of n units, accounting for decreasing proportions of the total variation among the units and corresponding to the eigenvalues (in decreasing order of size) in the correlation (or variance-covariance) matrix of the measurements. There are the same number of components as measurements, but the hope is that a small number of them will account for most of the variation and so reduce the dimensionality of a problem.

Cluster analysis aims, on the basis of p measurements on each, to group the units into sets that are "similar", again using the correlation matrix.

(b) (i) m400, m100 and m110h, the "short-run" variables, seem moderately well correlated. Most other correlations are moderate or low. m1500 does not appear to be noticeably correlated with any of the others; nor does high jump. Thus there seems to be only one obvious cluster.

(ii) There are 4 eigenvalues greater than 1, and they take up 69% of the variation, so this is a reasonable cut-off point to take as a basis for interpretation. Note that short times and long distances show good performance.



(iii) PC1 is a general average measure of good performance, bearing in mind the remark above. [The first PC of a correlation matrix very often is this.] m100, long jump, m400, m110h, discus and javelin are the major contributors to it.

PC2 is a contrast between high jump and m400.

PC3 is largely m1500, with less contribution from others.

PC4 is pole vault, with some contribution from other jumping measures.

(iv) Raw Euclidean distances give undue weight to variables with a longer time or distance characteristic. It would be better to scale the measures so that each had variance 1. If some were thought more important than others, they could be given greater weighting in the calculation.

(v) The dendrogram gives closest similarity to 15 and 16; this agrees with the plot of PC3 against PC4, but not PC1 and PC2.

Also 18 and 25 are very 'similar' in the cluster analysis, but 25 is rather an outlier on the PC plots. 30 is the last to come in to the clusters; this is similar to PC3/PC4 but not PC1/PC2. On the other hand, 31 joins a cluster quite soon but is an outlier on both PC plots. 10, 22, 30 form a distinctive cluster, they are also separated from most other points on the PC3/PC4 plot but not on PC1/PC2.

In summary, the dendrogram and PC3/PC4 plot give some of the same information; but PC1/PC2 matches neither.