THE ROYAL STATISTICAL SOCIETY

2006 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.

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, Applied Statistics, Paper I, 2006. Question 1

(i) The AR(p) time series model is

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$

where Y_t represents the series at time t and ϕ_0, ϕ_1, …, ϕ_p are constants. The $\{\varepsilon_t\}$ are "pure error" or "white noise" random terms, independently identically distributed N(0, σ_ε^2), and not correlated with $\{Y_t\}$.

The MA(q) time series model is

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

where $\{Y_t\}$, $\{\varepsilon_t\}$ are as above and θ_1, θ_2, …, θ_q are constants.

(ii)(a) $Y_t = 5 + \varepsilon_t - 0.3\varepsilon_{t-1}$ is a stationary process. $E[\varepsilon_t] = 0$ (as in part (i)), so

$$E[Y_t] = 5 + E[\varepsilon_t] - 0.3E[\varepsilon_{t-1}] = 5.$$

Also, again using the conditions in part (i),

$$\mathrm{Var}(Y_t) = 0 + \mathrm{Var}(\varepsilon_t) + (0.3)^2 \mathrm{Var}(\varepsilon_{t-1}) = 1.09\sigma_\varepsilon^2.$$

The autocovariance is

$$\gamma_k = \mathrm{Cov}(Y_t, Y_{t-k}) = \mathrm{Cov}(5 + \varepsilon_t - 0.3\varepsilon_{t-1},\ 5 + \varepsilon_{t-k} - 0.3\varepsilon_{t-k-1})$$

$$= \begin{cases} -0.3\mathrm{Var}(\varepsilon_{t-1}) = -0.3\sigma_\varepsilon^2 & \text{if } k = 1 \\ -0.3\mathrm{Var}(\varepsilon_t) = -0.3\sigma_\varepsilon^2 & \text{if } k = -1 \\ 0 & \text{otherwise} \end{cases}$$

Thus the autocorrelation ρ_k is $-\dfrac{0.3}{1.09} = -0.275$ for $k = \pm 1$, and 0 otherwise.

So the partial autocorrelation function decays to 0 and is negative.

Solution continued on next page

(ii)(b) $Y_t = 34 - 0.1Y_{t-2} + \varepsilon_t$ is stationary (this is given in the question).

$$\therefore E[Y_t] = 34 - 0.1E[Y_t], \quad \text{so } E\left[Y_t\right] = \frac{34}{1+0.1} = 30.91.$$

Also,

$$\text{Var}\left(Y_t\right) = \left(0.1\right)^2 \text{Var}\left(Y_t\right) + \sigma_\varepsilon^2, \quad \text{so Var}\left(Y_t\right) = \frac{\sigma_\varepsilon^2}{1-(0.1)^2} = 1.01\sigma_\varepsilon^2.$$

For the autocovariance, consider

$$\text{Cov}\left(Y_t, Y_{t-k}\right) = \text{Cov}\left(Y_t,\ 34 - 0.1Y_{t-k-2} + \varepsilon_{t-k}\right) = -0.1\ \text{cov}\left(Y_t, Y_{t-k-2}\right).$$

$\therefore \gamma_k = -0.1\gamma_{k-2}$ and $\rho_k = -0.1\rho_{k-2}$.

However, $\rho_1 = 0$ and $\rho_2 = -0.1$. Hence we have $\rho_3 = \rho_5 = \rho_7 = \ldots = 0$, and also $\rho_2 = -0.1$, $\rho_4 = (0.1)^2$, $\rho_6 = -(0.1)^3$, $\rho_8 = (0.1)^4$, … .

The partial autocorrelation function cuts off sharply after $k = 2$ (with a negative spike at $k = 2$).

(iii) $Y_t = 54 - 0.2Y_{t-1} + \varepsilon_t$.

$\therefore \left(1 + 0.2L\right)Y_t = 54 + \varepsilon_t$ where L represents the "backward shift" operator.

$$\therefore Y_t = \frac{54+\varepsilon_t}{\left(1+0.2L\right)} = \left(1 - 0.2L + \left(0.2\right)^2 L^2 - \left(0.3\right)^3 L^3 + \ldots\right)\left(54 + \varepsilon_t\right)$$

$$= 54 + \varepsilon_t - 0.2\varepsilon_{t-1} + \left(0.2\right)^2 \varepsilon_{t-2} - \left(0.2\right)^3 \varepsilon_{t-3} + \ldots .$$

Graduate Diploma, Applied Statistics, Paper I, 2006. Question 2

(i) (a) Variance and covariance depend on the units of measurement, but correlation does not. These variables are in completely different units of measurement, and this would seriously influence the results if the covariance matrix were used. There is also the issue (see (ii)(a)) of whether, for example, length of road should be measured in miles or kilometres. It would not be appropriate to use the covariance matrix.

(b) When the principal components are arranged in order of size of eigenvalues, as here, "cumulative proportions" are found by adding eigenvalues from the left, ending with a total (in this example) of 4, the number of variables. The "proportions" for each eigenvalue here are

$$\frac{2.23}{4}\times 100 = 55.75\%, \quad \frac{1.33}{4}\times 100 = 33.25, \quad \frac{0.25}{4}\times 100 = 6.25\%,$$

so the cumulative proportions are 55.75, 89.00, 95.25 and 100%.

(c) Most of the variation is explained by PC1 and PC2, leaving only 11% for PC3 and PC4 together. An explanation of the data might therefore be given in terms of these two. The coefficients in the components are the weightings of the four variables in each one.

PC1 is a (weighted) average of all four variables, as often happens when using the correlation matrix; population is of less importance than the other three variables.

PC2 is mainly a contrast between population and length of road. This might reflect the effect of length of road outside main population centres.

PC3 seems to be a contrast between number of drivers and (population density, length of road) but does not contribute much to the total. It might be some form of measure of the number of drivers outside main population centres.

PC4 is a contrast between fuel consumption and the others, but contributes little. Perhaps it is mainly a contrast between number of drivers and fuel consumption, and thus could give a measure of the number of drivers with high fuel consumption.

(Note that components that contribute little and might reasonably be ignored can sometimes give information about what is not important.)

Solution continued on next page

(ii) (a) Since the correlation matrix has been used, changing the units as suggested would make no difference to the results. Identical principal components and regression analyses would be obtained. The whole analysis is independent of the original scales of measurement.

(b) A simple explanation, with fairly high R^2, comes from PC1, and shows that deaths are positively related to number of drivers, length of road and fuel consumption, but depend little on population density. Adding PC3 and PC4 improves the explanation and leads to a very high R^2. Note that the interpretations depend on the signs of the coefficients.

(iii) Unless there are correlations among the predictor variables (which of course there often are), there is nothing to gain from using principal components. The principal components themselves are uncorrelated, so this makes model selection easier if a "suitable" subset is known. However, the principal components are not always easy to interpret as they are purely mathematical constructs, and thus the resulting regression models may not be easy to interpret. (Note for example that the negative coefficient of PC4 in the regression analysis in part (ii) suggests that the number of deaths is positively related to drivers with low fuel consumption, which seems strange.)

It can be difficult to choose a "suitable" subset of principal components. In this example, PC2 appeared important in the explanation of the data in part (i) but did not enter the regression analysis in part (ii). On the other hand, it can happen that components with low eigenvalues are important in the regression – as appears to have happened with PC3 and PC4 here. One way to try to deal with this is to include the response variable in the principal components analysis and select PCs with high coefficients of the response variable and large eigenvalues; but there is no guarantee that important regression variables will be identified.

Graduate Diploma, Applied Statistics, Paper I, 2006. Question 3

(i) Both methods can be carried out on data from groups of items for which there are multiple measurements. In cluster analysis the groups are not known already, but are constructed from the data; the analysis investigates possible groupings, based on the multiple measurements. Discriminant analysis is carried out when the (two or more) groups are known already and the aim is to find the (linear) combination of the variables which best separates the items into their groups. The method assumes that the data in each group are from multivariate Normal distributions with similar variance-covariance structures.

As an example, discriminant analysis would be appropriate for distinguishing between those animals which survive in a hard environment and those which do not, using a range of physiological measurements on bodies. A suitable discriminatory combination of these measurements would have been found from several species for which it was known whether or not they survive. The same combination would then be applied to measurements on a different species to give a prediction of whether or not it would be likely to survive.

(ii) (a) For $d(x, y)$ to be a proper distance measure for X and Y, we require: $d(x, x) = 0$; $d(x, y) = d(y, x)$; $d(x, y) \leq d(x, z) + d(z, y)$ for all Z.

Certainly $d(A, A) = d(B, B) = d(C, C) = 0$, from the diagonal entries.

Also, $d(A, B) = 3.9 = d(B, A)$; and similarly for $d(A, C)$ and $d(B, C)$.

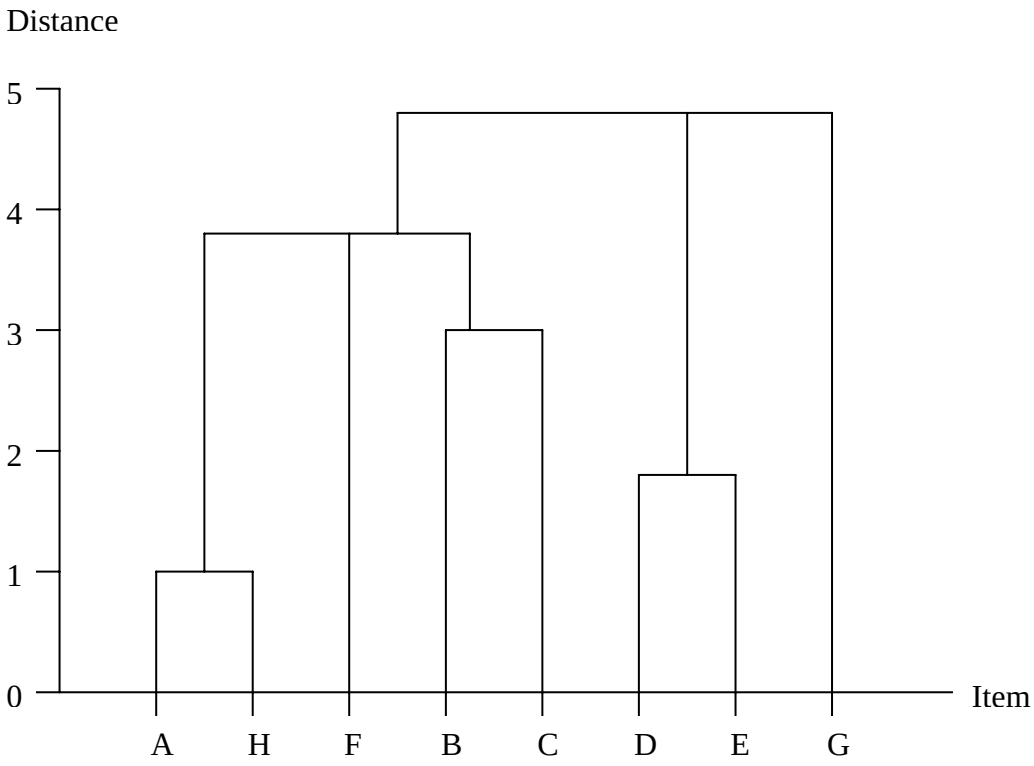
Finally, $d(A, C) = 5.5$, with $d(A, B) + d(B, C) = 3.9 + 3.0 = 6.9 > 5.5$; $d(B, C) = 3.0$, with $d(B, A) + d(A, C) = 3.9 + 5.5 = 9.4 > 3.0$; and $d(A, B) = 3.9$ with $d(A, C) + d(C, B) = 5.5 + 3.0 = 8.5 > 3.9$.

(b) At stage 1, items A and H have distance 1.0 and these form a cluster. Cluster 2 is D and E with distance 1.7, so stage 2 has two clusters, AH and DE. The next cluster is BC with distance 3.0. The next distance is 3.7, for AF and for BH; as shown in the dendrogram, we therefore combine ABCFH as a cluster at this distance, from which the cluster DE is still separate. The next distance is 4.6, for CD and CG; thus at this distance we are down to a single cluster.

[Note. In any practical case we may specify the number of clusters required as a reasonable summary of the data or we may specify the maximum distance for combining clusters. As an illustration of the latter, suppose we specified distance 3.1; at this distance there are 3 clusters, AH, BC and DE, and the other items stand on their own.]

(c) The dendrogram appears rather "straggly". A and H seem closely similar, as do D and E, but otherwise there are no very clear and concise clusters. There is not really any strong evidence of "distinct categories" by the single linkage cluster analysis used here, though different methods of clustering might give different structures.

The dendrogram is on the next page


Distance
5
4
3
2
1
0
Item
A
H
F
B
C
D
E
G

(i) (a) The probability function of the binomial distribution B(m, π) written in exponential form is

$$f(y,\pi) = \exp\left[\log\binom{m}{y} + y\log\pi + m\log(1-\pi) - y\log(1-\pi) \right]$$

so $$\log f(y_i,\pi_i) = y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) + \text{constant} \ .$$

As y_i is on the right of this equation, it is in canonical form, and the multiplier of y_i is the natural parameter, which is therefore

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

[This is the log-odds, or logit.] The generalised linear model sets this link function equal to the linear predictor.

(b) The *odds* is the ratio of probabilities of "success" and "failure" for Y_i, i.e. $\pi_i/(1 - \pi_i)$. The *log-odds* is simply the logarithm (base e) of this, as used in the link function.

After the generalised linear model has been fitted, the (estimated) value of η_i is obtained – this is the estimate of the log-odds. We have

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

so the estimate of the odds is $\exp(\eta_i)$.

Given the necessary standard errors (SE), approximate 95% confidence limits are "estimate ± 1.96×SE" for each of odds and log-odds. The details are in (ii)(c) below.

(ii) (a) We are not told how the sampling was carried out, so the independence of observations is not guaranteed; neither is the randomness.

The analysis would be appropriate if a random sample of data from a larger population has been selected, omitting multiple births (twins etc) and only using a mother once if she has had more than one child at different times [to avoid probable lack of independence]. Many hospitals would need to be represented in the sampling, as well as home births. It would not be appropriate to use this analysis if the "group of women" mentioned came from a limited area, for example by studying all births from the local hospital over a few years.

Solution continued on next page

(b) Step 1 chooses the single predictor variable which reduces the scaled deviance as much as possible from the "constant only" model. Clearly this is GEST, the length of gestation period, which reduces the deviance by 339.37 (on 1 df). We next consider adding AGE, and this step further reduces the deviance by 6.566, also on 1 df; this is significant as an observation from χ^2 with 1 df, so AGE should be included. So we use AGE and GEST in the model.

(c) The coding AGE = 0, GEST = 0 gives

$$\hat{\eta} = -1.7659, \quad \mathrm{SE}(\hat{\eta}) = 0.1296.$$

The estimate of the odds is exp(–1.7659) = 0.171.

95% confidence limits for η are –1.7659 ± 1.96×0.1296, i.e. (–2.020, –1.512), so the corresponding limits for the odds are (0.1327, 0.2205) after exponentiating.

The estimate of the probability of mortality is

$$\hat{\pi} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} = \frac{0.171}{1.171} = 0.146.$$

Similarly, the upper and lower limits of the 95% confidence interval for this probability are as follows.

$$\text{Lower limit: } \frac{0.1327}{1.1327} = 0.117; \quad \text{upper limit: } \frac{0.2205}{1.2205} = 0.181.$$

(d) GEST is now to be coded as 1, AGE remaining zero. The log-odds ratio for this group compared to the group in (c) is thus simply the value of the GEST parameter, i.e. –3.2886.

So the odds ratio is exp(–3.2886) = 0.0373.

95% confidence limits for this log-odds ratio are –3.2886 ± 1.96×0.1846, i.e. (–3.650, –2.927). Thus the limits for the odds ratio are exp(–3.650) = 0.026 and exp(–2.927) = 0.054.

Graduate Diploma, Applied Statistics, Paper I, 2006. Question 5

(i)(a) The appropriate model for log(y) is $\log y_i = \log a + bx_i + \varepsilon_i'$, where the $\{\varepsilon_i'\}$ are independent identically Normally distributed errors, N(0, σ^2).

Thus for the untransformed data we have $y_i = ae^{bx_i}\varepsilon_i$ where the ε_i are lognormal.

(i)(b) Denote log(y) by Y and log(a) by A so that $Y_i = A + bx_i + \varepsilon_i'$. The normal equations are obtained by minimising

$$S = \sum_i \left(Y_i - A - bx_i\right)^2 .$$

Differentiating with respect to A and B leads to the following (strictly speaking it should be checked that these are *minimising* values).

$$\frac{\partial S}{\partial A} = -2\sum_i \left(Y_i - A - bx_i\right).$$

Setting this equal to 0 gives $\bar{Y} = \hat{A} + \hat{b}\bar{x}$ as one normal equation. So $\hat{A} = \bar{Y} - \hat{b}\bar{x}$.

$$\frac{\partial S}{\partial b} = -2\sum_i x_i \left(Y_i - A - bx_i\right).$$

Setting this equal to 0 gives $\sum_i x_i Y_i = \hat{A}\sum_i x_i + \hat{b}\sum_i x_i^2$.

Thus $\sum x_i Y_i = \left(\bar{Y} - \hat{b}\bar{x}\right)\sum x_i + \hat{b}\sum x_i^2$

$$\therefore \hat{b} = \frac{\Sigma x_i Y_i - \dfrac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^{\,2} - \dfrac{(\Sigma x_i)^2}{n}} \quad \left[\text{or } \frac{n\Sigma x_i Y_i - (\Sigma x_i)(\Sigma Y_i)}{n\Sigma x_i^{\,2} - (\Sigma x_i)^2}\right]$$

Solution continued on next page

(ii) Here we similarly minimise $S = \sum_i \left(y_i - ae^{bx_i} \right)^2$, giving two normal equations as follows.

$$\frac{\partial S}{\partial a} = -2\sum_i \left(y_i - ae^{bx_i} \right) e^{bx_i} .$$

Setting this equal to 0 gives $\sum_i y_i e^{\hat{b}x_i} = \hat{a}\sum_i e^{2\hat{b}x_i}$.

$$\frac{\partial s}{\partial b} = -2\sum_i \left(y_i - ae^{bx_i} \right) ax_i e^{bx_i} .$$

Setting this equal to 0 gives

$$\hat{a}\sum x_i y_i e^{\hat{b}x_i} = \hat{a}^2 \sum x_i e^{2\hat{b}x_i} \quad \text{or} \quad \sum x_i y_i e^{\hat{b}x_i} = \hat{a}\sum x_i e^{2\hat{b}x_i} .$$

From the first normal equation we have $\hat{a} = \dfrac{\sum_i y_i e^{\hat{b}x_i}}{\sum_i e^{2\hat{b}x_i}}$.

Inserting this in the second gives

$$\left(\sum_i x_i y_i e^{\hat{b}x_i} \right) \left(\sum_i e^{2\hat{b}x_i} \right) = \left(\sum_i y_i e^{\hat{b}x_i} \right) \left(\sum_i x_i e^{2\hat{b}x_i} \right)$$

(This would require an iterative method of solution such as Newton-Raphson.)

(iii) (a) First model: $\hat{a} = e^{0.918} = 2.504$; $\hat{b} = 1.19$.

Second model: $\hat{a} = 2.45$; $\hat{b} = 1.20$.

The estimates are very similar.

(b) The situation is not clear-cut. The standardised residuals from the first model are perhaps more scattered than those from the second, and with an appearance of becoming less variable as *x* increases. For the second model, there is only one outlier among the residuals but the pattern may be slightly skew. The outlier in this model corresponds to the point where there is a slight jump in the original scatter diagram, and this also produces a somewhat high standardised residual from the first model. Perhaps there is a very slight preference for the second model.

(c) Any existing knowledge about the system will be helpful, as will any prior information on the error structure – is it Normal or lognormal (or neither)?

(i) The single variable which reduces the residual sum of squares most from the model with intercept alone is x_4, so forward selection first introduces this variable. It then examines adding another variable to x_4. The smallest residual sum of squares for x_4 and one other variable is when x_1 is taken with x_4. Next to be entered would be x_2 but this does not seem to make much difference compared with (x_1, x_4). Adding x_3 as well does not seem to make a worthwhile difference either. Formal tests for significance at each stage are as follows.

(1) Adding x_4. The SS for adding x_4 is 2715.764 – 883.867 = 1831.897, and the remaining residual then has 11 df. So the test statistic is

$$\frac{\frac{1831.897}{1}}{\frac{883.867}{11}} = 22.80,$$

which is very highly significant as an observation from $F_{1,11}$; there is very strong evidence that x_4 should be included in the model.

(2) Adding x_1. The SS for doing this is 883.867 – 74.762 = 809.105, and the remaining residual has 10 df. The test statistic is

$$\frac{\frac{809.105}{1}}{\frac{74.762}{10}} = 108.22,$$

which is very highly significant as an observation from $F_{1,10}$; there is very strong evidence that x_1 should also be included in the model.

(3) Adding x_2. The SS for doing this is 74.762 – 47.973 = 26.789, and the remaining residual has 9 df. The test statistic is

$$\frac{\frac{26.789}{1}}{\frac{47.973}{9}} = 5.03,$$

which is not (quite) significant at the 5% level as an observation from $F_{1,9}$ (the 5% critical point is 5.12). Judged at the 5% level, there is no evidence that x_2 should also be included in the model.

Thus the model reached by forward selection has x_1 and x_4.

Solution continued on next page

(ii) Not if there are correlations among the predictor variables. Starting from the full model and omitting variables can give very different results from forward selection.

(iii) Mallows' C_p is a diagnostic statistic designed to help in identifying a "best subset" model. The definition of C_p is

$$C_p = \frac{RSS_p}{s^2} - (n - 2p)$$

where

n is the number of data values,

p is the number of parameters (including the constant) in the model being investigated,

RSS_p is the residual sum of squares for the model being investigated,

s^2 is the residual mean square for the full model using all possible predictors.

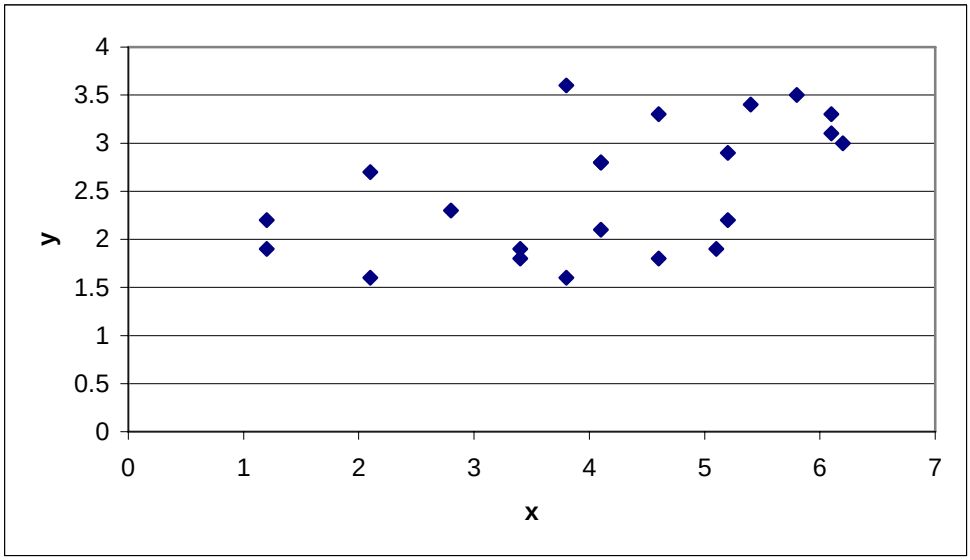
It can be shown that, for a satisfactory model, $E[C_p] = p$. Looking at the values of C_p given in the question, this suggests that the (x_1, x_4) model may be not quite "good enough" as it has a C_p value of 5.50 as opposed to an expectation of 2. All of the three-predictor models seem satisfactory apart from (x_2, x_3, x_4).

(iv) The two-predictor model (x_1, x_2) appears good, both from its residual sum of squares and from its C_p value. It also has a higher R^2 value than any of the other models with fewer than three predictors. This model has not appeared in the forward selection because neither x_1 nor x_2 was the first variable chosen, but it appears to be the best parsimonious model.

(v) It is always important to have information about the practical situation, to avoid proposing a model that researchers or practitioners in the field consider unreasonable. They may do so from their general technical knowledge of the field and/or in the light of earlier work which has shown that some potential variables in fact have little relation to the response y. Correlations among the x variables may, mathematically, lead to such "unreasonable" models. If there is a choice, those variables that are easier and cheaper to measure accurately would often be preferred. Why have the variables x_1, x_2, x_3, x_4 in the example been suggested?

Graduate Diploma, Applied Statistics, Paper I, 2006. Question 7

(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

$$Y_{ij} = a + bx_i + \varepsilon_{ij} \qquad i = 1, 2, \ldots, 13$$
$$j = 1 \text{ or } 2 \text{ or } 3, \text{ depending on the value of } i.$$

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, Applied Statistics, Paper I, 2006. Question 8

Part (i)

Totals are as follows

For A $A1$: 56; $A2$: 78; $A3$: 83.

For B $B1$: 145; $B2$: 72.

Grand total: 217.

Sum of squares of observations: 1977.

The (corrected) total sum of squares is $1977 - \frac{217^2}{30} = 407.367$, with 29 df.

The sum of squares for factor A is

$$\frac{56^2}{10} + \frac{78^2}{10} + \frac{83^2}{10} - \frac{217^2}{30} = 41.267, \text{ with 2 df.}$$

The sum of squares for factor B is

$$\frac{145^2}{15} + \frac{72^2}{15} - \frac{217^2}{30} = 177.633, \text{ with 1 df.}$$

The sum of squares for the interaction AB is

$$\frac{32^2}{5} + \frac{24^2}{5} + \frac{50^2}{5} + \frac{28^2}{5} + \frac{63^2}{5} + \frac{20^2}{5} - \frac{217^2}{30} - \text{SS for } A - \text{SS for } B = 62.067,$$

with 2 df.

The residual SS is obtained by subtraction. This has 29 – 2 – 1 – 2 = 24 df.

A is a fixed factor in both (a) and (b) below; its interpretation is the same in the two parts. B is fixed in (a) and random in (b), so its interpretation is different in the two parts, and this also applies for the interaction AB.

Solution continued on next page

(a) The model is

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

$i = 1, 2, 3;\ \ j = 1, 2;\ \ k = 1, 2, \ldots, 5$

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_i (\alpha\beta)_{ij} = 0, \quad \sum_j (\alpha\beta)_{ij} = 0\,.$$

Here μ, α_i, β_j, $(\alpha\beta)_{ij}$ are all unknown constants, representing the grand mean, the effect of level i of A, the effect of level j of B, and the interaction between A at level i and B at level j, respectively The "error" terms $\{\varepsilon_{ijk}\}$ are independent Normal random variables, N(0, σ^2), with σ^2 constant.

The analysis of variance is as follows. A column of expected mean squares ($E[MS]$) is inserted in the table.

Source of variation	df	Sum of squares	Mean square	$E[MS]$	F value
A	2	41.267	20.633	$\sigma^2 + \left(\frac{2\times5}{2}\right)\Sigma\alpha_i^{\,2}$	20.633/5.267 = 3.92
B	1	177.633	177.633	$\sigma^2 + \left(\frac{3\times5}{1}\right)\Sigma\beta_j^{\,2}$	177.633/5.267 = 33.73
AB	2	62.067	31.033	$\sigma^2 + \left(\frac{5}{2\times1}\right)\Sigma\Sigma(\alpha\beta)_{ij}^{\,2}$	31.033/5.267 = 5.89
Residual	24	126.400	5.267		$= \hat{\sigma}^2$
Total	29	407.367			

Tests for the null hypotheses "all $\alpha_i = 0$", "all $\beta_j = 0$", "all $(\alpha\beta)_{ij} = 0$" use the given F values. The upper 5% point of $F_{2,24}$ is 3.40 and the upper 1% point is 5.61, so the first of these is significant at the 5% level and the third at the 1% level. For $F_{1,24}$, the upper 0.1% point is 14.03, so the second is very highly significant. There is evidence that both main effects and the interaction are important.

As the interaction is significant, the interpretation should be based on a 2-way table of AB means. The response for $B2$ is much the same at each level of A; however, for $B1$ the response at $A1$ is well below the other two, with $A3$ a little larger than $A2$.

Solution continued on next page

(b) The model is

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

$i = 1, 2, 3;\ \ j = 1, 2;\ \ k = 1, 2, \ldots, 5.$

Here μ and the α_i are unknown constants, with $\Sigma\alpha_i = 0$, as before, and the "error" terms $\{\varepsilon_{ijk}\}$ are independent Normal random variables, N(0, σ^2), with σ^2 constant, as before.

The β_j are uncorrelated random variables, uncorrelated with the $(\alpha\beta)_{ij}$ and with the $\{\varepsilon_{ijk}\}$, with mean 0 and constant variance σ_B^2.

The $(\alpha\beta)_{ij}$ represent the interaction. For each level of A (i.e. for each i) they are uncorrelated random variables, uncorrelated with the β_j and with the $\{\varepsilon_{ijk}\}$, with mean 0 and constant variance σ_{AB}^2. For each level of B (i.e. for each j) they are constants with $\Sigma_i(\alpha\beta)_{ij} = 0$.

Assumptions of Normality for the β_j and $(\alpha\beta)_{ij}$ are added for the formal inferences, so that these become independent Normal random variables.

The analysis of variance is as follows. A column of expected mean squares (E[MS]) is again inserted in the table.

Source of variation	df	Sum of squares	Mean square	E[MS]	F value
A	2	41.267	20.633	$\sigma^2 + 5\sigma_{AB}^2 + \left(\frac{2\times5}{2}\right)\Sigma\alpha_i^2$	20.633/31.033 = 0.665
B	1	177.633	177.633	$\sigma^2 + (5\times3)\sigma_B^2$	177.633/5.267 = 33.73
AB	2	62.067	31.033	$\sigma^2 + 5\sigma_{AB}^2$	31.033/5.267 = 5.89
Residual	24	126.400	5.267	σ^2	
Total	29	407.367			

The null hypotheses are "all $\alpha_i = 0$", "$\sigma_B^2 = 0$" and "$\sigma_{AB}^2 = 0$". The expected mean squares indicate how these are tested.

Solution continued on next page

The F value to test the first null hypothesis is 0.665, which is referred to $F_{2,2}$. This is not significant; there is no evidence against this null hypothesis.

The F value to test the second is 33.73, which is referred to $F_{1,24}$. This is very highly significant (see part (a) above); there is very strong evidence against this null hypothesis.

The F value to test the third is 5.89, which is referred to $F_{2,24}$. This is significant at the 5% level (see part (a) above); there is evidence against this null hypothesis.

Estimates of σ_B^2 and σ_{AB}^2 can be obtained if required.

The explanation should be in terms of there being evidence for variation among the effects of the levels of B in the underlying population of all such levels, and similarly for the AB interaction.

Part (ii)

Suppose that $A1$, $A2$, $A3$ are three alternative cultivation treatments in an agricultural trial, while $B1$, $B2$ are two sites upon which that experiment is carried out. If $B1$, $B2$ are the only two available sites about which inferences are to be made, (a) is appropriate. If $B1$, $B2$ are selected (at random) from several available sites and inference is to be made for the whole collection of sites, (b) is appropriate.