THE ROYAL STATISTICAL SOCIETY

2007 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, 2007. Question 1

(i) AR(1): $Y_t = \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t$

MA(1): $Y_t = \phi_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1}$

where the $\{\phi_i\}$ are constants and the $\{\varepsilon_t\}$ are uncorrelated identically distributed random variables which are also uncorrelated with the $\{Y_t\}$.

(ii) $Y_t = \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t$

$= \phi_0 + \phi_1 \left(\phi_0 + \phi_1 Y_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t$

$= \phi_0 + \phi_1 \left[\phi_0 + \phi_1 \left(\phi_0 + \phi_1 Y_{t-3} + \varepsilon_{t-2}\right) + \varepsilon_{t-1}\right] + \varepsilon_t$

$= \phi_0 \left(1 + \phi_1 + \phi_1^2\right) + \phi_1^3 Y_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t$

$= \ldots$

$= \phi_0 \left(1 + \phi_1 + \phi_1^2 + \ldots\right) + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + \ldots$

$= \dfrac{\phi_0}{1-\phi_1} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \ldots$.

(iii) We let σ_ε^2 denote the common variance of the $\{\varepsilon_t\}$.

For AR(1), we have

$\mathrm{Cov}\left(Y_t, Y_{t-r}\right)$

$= \mathrm{Cov}\left(\dfrac{\phi_0}{1-\phi_1} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \ldots,\ \dfrac{\phi_0}{1-\phi_1} + \varepsilon_{t-r} + \phi_1 \varepsilon_{t-r-1} + \phi_1^2 \varepsilon_{t-r-2} + \ldots \right)$

[using (ii) above].

Solution continued on next page

Putting $r = 0$ gives $\mathrm{Var}(Y_t) = \sigma_\varepsilon^2 \left(1 + \phi_1^2 + \phi_1^4 + ...\right) = \dfrac{\sigma_\varepsilon^2}{1 - \phi_1^2}$.

[Note: this also follows directly from the expression at the end of part (ii), as the $\{\varepsilon_t\}$ are uncorrelated.]

Putting $r = 1$ gives $\mathrm{Cov}(Y_t, Y_{t-1}) = \sigma_\varepsilon^2 \phi_1 \left(1 + \phi_1^2 + \phi_1^4 + ...\right) = \dfrac{\sigma_\varepsilon^2 \phi_1}{1 - \phi_1^2}$.

Putting $r = 2$ gives $\mathrm{Cov}(Y_t, Y_{t-2}) = \sigma_\varepsilon^2 \phi_1^2 \left(1 + \phi_1^2 + \phi_1^4 + ...\right) = \dfrac{\sigma_\varepsilon^2 \phi_1^2}{1 - \phi_1^2}$.

In general, we get $\mathrm{cov}(Y_t, Y_{t-k}) = \dfrac{\sigma_\varepsilon^2 \phi_1^k}{1 - \phi_1^2}$.

Thus the autocorrelation function is ϕ_1^k.

The partial autocorrelation function is a decaying function.

For MA(1), we have $\mathrm{Var}(Y_t) = \mathrm{Var}(\phi_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1}) = (1 + \phi_1^2)\sigma_\varepsilon^2$, as the $\{\varepsilon_t\}$ are uncorrelated.

$$\mathrm{Cov}(Y_t, Y_{t-r}) = \mathrm{Cov}(\phi_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1},\ \phi_0 + \varepsilon_{t-r} + \phi_1 \varepsilon_{t-r-1})$$

$$= \begin{cases} \sigma_\varepsilon^2 \phi_1 & \text{for } r = 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus the autocorrelation function is $\dfrac{\phi_1}{1 + \phi_1^2}$ for $r = 1$, and 0 otherwise.

The partial autocorrelation function has value 0 for $r \geq 2$.

Solution continued on next page

(iv)(a) This is ARMA(1), a mixture of AR(1) and MA(1).

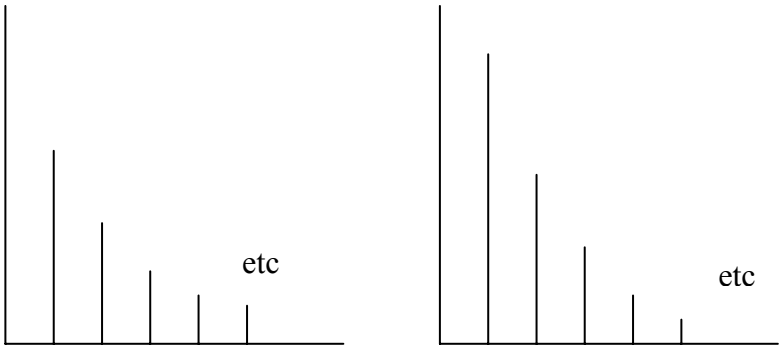
(b) $E(Y_t) = 20 + 0.8E(Y_{t-1})$ and so by stationarity we have $E(Y_t) = 20 + 0.8E(Y_t)$, so $E(Y_t) = 20/0.2 = 100$.

Again using stationarity, we have

$$\mathrm{Var}(Y_t) = 0 + (0.8)^2 \mathrm{Var}(Y_t) + \sigma_\varepsilon^2 (1 + 0.04 - (2 \times 0.8 \times 0.2))$$

which gives $\mathrm{Var}(Y_t) = 2\sigma_\varepsilon^2$.

(c) The graphs are only intended as rough sketches. The PACF decays more quickly than the ACF.



ACF PACF

(d) The ACF and PACF may suggest a simple AR(1) or MA(1) model. If not, possibilities are AR(m) or MA(n) or a mixture ARMA(m, n). It is usually sensible to begin with as simple a model as possible and fit it, then examine its goodness of fit and the residuals. If necessary, more complicated models should be tried.

Graduate Diploma, Applied Statistics, Paper I, 2007. Question 2

(i) Often, as in the case in part (ii), a set of data will show several substantial correlations between the p variables in the study. Principal component analysis aims to explain the relations among these p variables in terms of fewer than p components which are uncorrelated linear combinations of them.

(ii) (a) All correlations between pairs of variables are positive. There are very high correlations between POP and EMPLOY and between SCHOOL and HOUSE. There are quite high correlations between SCHOOL and SERVICES and between SERVICES and HOUSE. Of the remaining correlations, some are moderate and some very low.

A possible subset of strongly correlated variables would be POP and EMPLOY; another would be SCHOOL, SERVICES and HOUSE.

(b) Principal components analysis using a covariance matrix is heavily influenced by the units in which the variables have been measured. Here the measurement scales are quite different; it is essentially meaningless to combine the original variables. Use of the correlation matrix corrects for this by "standardising" each variable initially.

(iii) (a) These are the eigenvalues and corresponding eigenvectors of the correlation matrix $\mathbf{S}$, found in the usual way by solving $|\mathbf{S} - \lambda\mathbf{I}| = 0$ and then, for each eigenvalue λ, finding the eigenvector from $\mathbf{Sx} = \lambda\mathbf{x}$.

(b) The first principal component is (as is often the case) a weighted average of all the variables, with SERVICES having the largest coefficient and thus dominating the combination. It may be interpreted as measuring, in some sense, the overall numbers of people in the areas, the availability of professional services and prosperity in terms of house values. The second principal component is a contrast between (POP, EMPLOY) and the other three variables (SCHOOL, SERVICES, HOUSE), though SERVICES has only a small weighting. This may be interpreted as numbers of people versus some aspects of socio-economic status.

(c) Using the correlation matrix, the sum of all the eigenvalues is the number of variables, here 5. The sum of the eigenvectors for the two components that have been used is 2.8733 + 1.7966 = 4.6699. Thus these two components explain 4.6699/5 = 93.4% of the standardised variance.

(d) The first two principal components together leave very little of the total variation unexplained. There will be no other comparisons of noticeable size to be found. Because so little variation is left unexplained, the choice of only two principal components in this case seems justified.

Solution continued on next page

(e) In general this simple criterion, by itself, is not best, as it merely gives those whose contribution has been "above average" when using the correlation matrix. If, for example, all but one of the eigenvalues had been only slightly above 1, with the last being very small to balance, interpretation would have been relatively vague – none of the components would be "more important" than the others (apart from the last). There is also the point that the small eigenvalues and corresponding eigenvectors occasionally give useful information, pointing to redundant variables which need not be measured in future studies. The purpose of the study and the nature of the variables must always be kept in mind.

Graduate Diploma, Applied Statistics, Paper I, 2007. Question 3

(i) Group is coded simply as (0, 1). The three X variables are all continuous measurements, and a fitted model may well give some predictions outside the range [0, 1] altogether. The residuals are very unlikely to be Normally distributed or to have constant variance.

(ii) Discriminant analysis assumes that the data in each group follow a multivariate Normal distribution with the same variance-covariance matrix. (There is a degree of robustness to lack of Normality). It aims to find a linear function of the p measured variables ($x_1, x_2, \ldots, x_p$) such that the ratio (between groups variance) / (within groups variance) is maximised.

The method of finding the discriminant function, using $\mathbf{a} = \mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)$ where $\mathbf{S}$ is the pooled variance-covariance matrix, is illustrated in part (iii)(b) below.

An item should be assigned to group 1 if the x readings for it, inserted into the discriminant function, lead to a value nearer to $\bar{z}_1$ than to $\bar{z}_2$, where $\bar{z}_i$ is the value of the discriminant function at the mean values of the x variables in group i. (See part (v) below.)

Logistic regression fits a model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_i \beta_i x_i,$$

where p is the probability that an item is in group 1. This leads to

$$p = P\left(\text{group 1} \mid \mathbf{x}\right) = \frac{\exp\left(\beta_0 + \Sigma\beta_i x_i\right)}{1 + \exp\left(\beta_0 + \Sigma\beta_i x_i\right)}$$

and

$$P\left(\text{group 2} \mid \mathbf{x}\right) = \frac{1}{1 + \exp\left(\beta_0 + \Sigma\beta_i x_i\right)}$$

An item should be assigned to group 1 if the x readings for it, with the estimated values of the βs, lead to $\beta_0 + \Sigma\beta_i x_i > 0$. (See part (v) below.)

Fewer assumptions are required by this method, but the fit of the logistic regression model should be checked.

Solution continued on next page

(iii) (a) There are 16 degrees of freedom for the items in the variance-covariance matrix of group 1 and 14 for those of group 2. The matrix $\mathbf{S}$ is pooled from $\mathbf{S}_1$ and $\mathbf{S}_2$ as $\mathbf{S} = \frac{1}{30}\left(16\mathbf{S}_1 + 14\mathbf{S}_2\right)$.

The entry 20.12 for the pooled covariance between X1 and X2 is found as $\frac{1}{30}\left((16 \times 22.154) + (14 \times 17.794)\right)$.

(b) $\mathbf{a} = \mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)$

$$= \begin{pmatrix} 0.0278 & -0.0273 & -0.0047 \\ -0.0273 & 0.0959 & -0.0160 \\ -0.0047 & -0.0160 & 0.0300 \end{pmatrix} \begin{pmatrix} -10.91 \\ -6.643 \\ -7.15 \end{pmatrix} = \begin{pmatrix} -0.088 \\ -0.225 \\ -0.057 \end{pmatrix}$$

so the discriminant function is $-0.088x_1 - 0.225x_2 - 0.057x_3$.

(iv) The model fitted (with sensible rounding of the coefficient values) is

$$\log\left(\frac{p}{1-p}\right) = 47.700 - 0.102x_1 - 0.220x_2 - 0.097x_3 = a(\mathbf{x}), \text{ say.}$$

and we have $P(\text{group 1}) = \frac{e^{a(\mathbf{x})}}{1+e^{a(\mathbf{x})}}$ and $P(\text{group 2}) = \frac{1}{1+e^{a(\mathbf{x})}}$.

None of the coefficients of x-variables are significant. The constant is significantly different from 0 but its value has a very wide confidence interval. Models with a single x-variable might be examined to see if the full model is actually any better than a simpler one.

Solution continued on next page

(v) Discriminant analysis gives the following value for the discriminant function:

$$(-0.088\times 175)-(0.225\times 71)-(0.057\times 133)=-38.956.$$

To find whether this is nearer to $\bar{z}_1$ than to $\bar{z}_2$ (see part (ii) above), first find their mean $\frac{1}{2}(\bar{z}_1+\bar{z}_2)$

$$=\frac{1}{2}\{-0.088(174.82+185.73)-0.225(69.824+76.467)-0.057(130.35+137.50)\}$$

$$=-39.956.$$

So assign to group 1 (since –38.956 > –39.956, and it is clear from the figures that $\bar{z}_1 > \bar{z}_2$).

Logistic Regression gives

$$a(\mathbf{x})=47.700-(0.102\times 175)-(0.220\times 71)-(0.100\times 133)=0.93>0,$$

so assign to Group 1.

(vi) There must be doubt whether the separate variance-covariance matrices are the same, so the logistic regression may be preferred – though a check of the fit of the logistical regression model should be undertaken.

Graduate Diploma, Applied Statistics, Paper I, 2007. Question 4

(i) We have that C_{ij} is distributed as Poisson(λ_i). There are N_i policyholders in cell i and the C_{ij} quantities for each of them are independent, so

$$C_i = \sum_{j=1}^{N_i} C_{ij} \sim \text{Poisson}(N_i\lambda_i)$$

and thus the mean of the number of claims in cell i is $N_i\lambda_i$.

(ii) A policyholder may make more than one claim, even if this is not very likely. Using the Poisson distribution allows for this, whereas the binomial does not, except as an approximation in a fairly large population with very small probability of repeat claims.

(iii) Writing C_{ij} as y_i, the Poisson model is $f(y_i) = \dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$.

Thus $\log f = -\lambda_i + y_i \log\lambda_i - \log(y_i!)$, so the natural link function is $\log\lambda_i$.

However, we observe the C_i as defined above, not the C_{ij}. So we can only fit a model to $\theta_i = N_i\lambda_i$, which gives $\log\theta_i = \log N_i + \log\lambda_i$; so $\log N_i$ is required as an offset for the model.

(iv) Following through the order as given in the question for terms to be added to the model, we examine the changes in scaled deviance. Each factor has 4 levels, so introducing the main effect of any new factor will use up 3 df, and correspondingly for interactions.

Upper 5% points are 7.815 for χ^2_3 and 12.592 for χ^2_6.

Model	*Change in scaled deviance*	*Change in df*	
1 → 2	238.16 – 225.57 = 12.59	3	Significant
2 → 3	225.57 – 137.39 = 88.18	3	(Very highly) significant
3 → 4	137.39 – 51.01 = 86.38	3	(Very highly) significant
4 → 5	51.01 – 40.06 = 10.95	6	Not significant

Similarly 5 → 6 and 6 → 7 are not significant changes.

Thus the most parsimonious model is 4, with only the three main effects. The residual df for this model will be 4^3 – 1 (for the constant) – 9 (for the main effects) = 54. Thus for this model we have (scaled deviance) ÷ df = 51.01/54 < 1, indicating that the model might be acceptable.

Solution continued on next page

(v) The parameter for DIST = 1 is 0. The parameter estimates for DIST = 2 and DIST = 3 do not differ significantly from 0. That for DIST = 4 is highly significantly different from zero; so this district seems to be very different from all the others. A possible recoding is therefore to code the fourth district as 1 and all the others as 0.

The AGE coefficients seem to decrease linearly. Thus age could be treated as a covariate with values 1, 2, 3 and 4, rather than as a factor.

Graduate Diploma, Applied Statistics, Paper I, 2007. Question 5

(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 mean square 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 (see question 2) or scatter diagrams (see question 6) 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, Applied Statistics, Paper I, 2007. Question 6

(i) Ozone is to be taken as y, the response variable. It is negatively correlated with wind, but the relation is probably curved; it is positively correlated with temperature, again curvilinear. Ozone and "rad" are related to some extent but the form of the relation is not at all clear.

The predictor variables show some correlations, particularly between temperature and wind, and possibly between temperature and "rad".

(ii) (a) An interaction is when the response to one predictor is affected by the level of other predictors present. It may be modelled as a product of predictors, as shown in part (ii)(b) below.

(b) A suitable model is

$$y = \beta_0 + \beta_1 t + \beta_2 w + \beta_3 r + \beta_{12}(tw) + \beta_{13}(tr) + \beta_{23}(wr) + \beta_{123}(twr) + \varepsilon$$

where t represents temperature, w wind and r "rad", the β terms are all constants and ε is a Normally distributed residual (error) term with mean 0 and constant variance σ^2. The terms with β_i represent main effects, those with β_{ij} represent two-variable interactions and that with β_{ijk} represents the three-variable interaction.

(iii) Examine interactions, starting with the highest (3-variable) one. The t value for this is not significant, so proceed to 2-variable interactions. Here, no 2-variable interactions appear significant either. So next examine the quadratic terms followed by the linear ones.

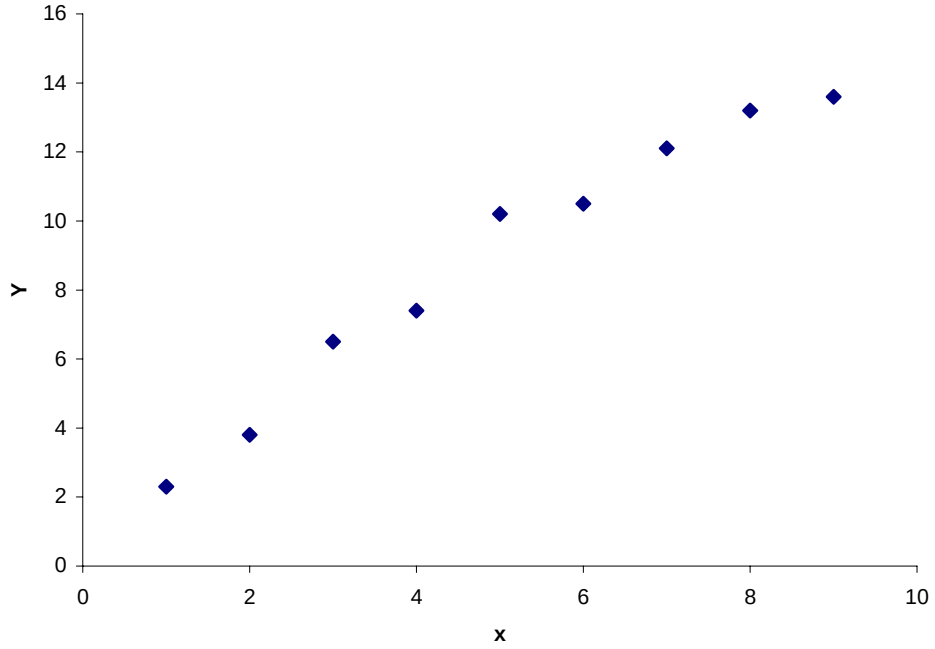
(iv) All of the terms in this model are statistically significant, the quadratic component of t appearing stronger than the linear part. R^2 is 71%, which is reasonably good. However, the plot of residuals against fitted values shows non-constant variance, with larger fitted values having more variable residuals, so the standard errors in the table, and any inferences based on them, are likely to be unreliable. The Normal plot is curved, not straight, suggesting non-Normality of the residuals. All in all, the validity of the fitted model is uncertain. Perhaps there are some particularly influential values, but the overall apparent poor fit of the model needs addressing first.

(v) Based on the pattern in the residuals, log y may sensibly be tried as a more satisfactory response variable.

The outliers suggested by the Cook's distance diagram should be examined to see what, if anything, is special or different about those data items. Consider whether any should be removed from the set of 111 data items (3 in 111 is not necessarily serious). Finally carry out the same checks for the model using log y as have been done for y.

Graduate Diploma, Applied Statistics, Paper I, 2007. Question 7

(i)



There is evidence of non-linearity, possibly consistent with two straight lines as suggested in the question (one for x from 1 to 5, the other for 5 to 9, the second having smaller gradient that the first). A possible alternative would be a quadratic in x, or $Y = \log x$.

(ii) The model fitted here is one with two straight lines with appropriate dummy variables X_1 and X_2:-

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \, .$$

The first of the given columns corresponds to the constant term β_0, which gives the expected value of Y when $X_1 = X_2 = 0$ (i.e. at $x = 5$). The second represents the values of the first dummy variable ($X_1 = -4, -3, -2, -1, 0, 0, 0, 0, 0$) with β_1 being the gradient of the first line segment, and similarly for the third column and the second line segment ($X_2 = 0, 0, 0, 0, 0, 1, 2, 3, 4,$). The model ensures that the lines intersect at $X_1 = X_2 = 0$ (i.e. at $x = 5$).

(iii) The model is the same but the coding is different: the new X_1 is simply the old X_1 plus 5. β_0 is now the intercept at $x = 0$. β_1 and β_2 are as in part (ii).

Solution continued on next page

(iv) We minimise $S = \sum_{1}^{9} \left(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 \right)^2$.

We do this by setting derivatives = 0.

$\dfrac{\delta S}{\delta \beta_0} = -2\Sigma \left(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 \right)$, so we have $0 = \Sigma Y - 9\hat{\beta}_0 - \hat{\beta}_1 \Sigma X_1 - \hat{\beta}_2 \Sigma X_2$.

$\dfrac{\delta S}{\delta \beta_1} = -2\Sigma X_1 \left(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 \right)$, so $0 = \Sigma X_1 Y - \hat{\beta}_0 \Sigma X_1 - \hat{\beta}_1 \Sigma X_1^{\,2} - \hat{\beta}_2 \Sigma X_1 X_2$.

$\dfrac{\delta S}{\delta \beta_2} = -2\Sigma X_2 \left(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 \right)$, so $0 = \Sigma X_2 Y - \hat{\beta}_0 \Sigma X_2 - \hat{\beta}_1 \Sigma X_1 X_2 - \hat{\beta}_2 \Sigma X_2^{\,2}$.

Thus the normal equations are $\begin{cases} 79.6 = 9\hat{\beta}_0 - 10\hat{\beta}_1 + 10\hat{\beta}_2 \\ -41.0 = -10\hat{\beta}_0 + 30\hat{\beta}_1 \\ 128.7 = 10\hat{\beta}_0 + 30\hat{\beta}_2 \end{cases}$

Solving these three simultaneous equations gives

$\hat{\beta}_0 = 9.871, \quad \hat{\beta}_1 = 1.924, \quad \hat{\beta}_2 = 1.000$.

(v) (a) One possibility, which would ideally need a few more data points, would be to fit a line to the first four points, another to the last four, and ignore the middle one. However, the present data suggest that $x = 5$ is part of the second line; there is perhaps more doubt as to whether it is part of the first.

With the present data set, the design matrix could be altered to have an intersection not at $x = 5$ but at a value either side of it to be found by trial and error.

(b) Model-fitting can be investigated statistically by looking at the residuals and through other diagnostics commonly given by computer programs. Ideally, though, choice of model should depend also on what is known about the nature of the data, i.e. the context of the real problem from which the data have arisen. For example, experience is that some medical data seem to contain "change points" where the direction of a linear relation changes. Also, use of logarithmic scales may be (at least) as good as fitting (say) a quadratic relationship, and such scales are often easier to explain in biological or medical contexts.

Graduate Diploma, Applied Statistics, Paper I, 2007. Question 8

(i) The model is as follows.

$y_{ijkl} = \mu$	Overall mean
$+ \alpha_i$	Fixed effect of ith treatment, $i = 1, 2$, with $\Sigma\alpha_i = 0$
$+ c_j$	Random effect of jth clinic, $j = 1, 2, \ldots, 8$
$+ d_{(j)k}$	Random effect of doctor k at clinic j, $k = 1, 2, 3$
$+ p_{(jk)l}$	Random effect of patient l of doctor k at clinic j, $l = 1, .., 4$
$+ \varepsilon_{ijkl}$	Residual

The residuals are independent N(0, σ^2) random variables.

The three random effects are independent Normal random variables with mean 0 and variances respectively σ_C^2, σ_D^2, σ_P^2.

These sets of random variables are all mutually independent.

Note that there are 96 patients, 24 doctors and 8 clinics.

(ii) The analysis of variance table 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	$E[MS]$
Between treatments	1	4240.04		
Between clinics within treatments	6	2599.49	433.248	$\sigma^2 + 4\sigma_P^2 + 12\sigma_D^2 + 48\sigma_C^2$
Between clinics	7	6839.53		
Between doctors within clinics	16	7429.58	464.349	$\sigma^2 + 4\sigma_P^2 + 12\sigma_D^2$
Between doctors	23	14269.11		
Between patients within doctors	72	25236.88	350.512	$\sigma^2 + 4\sigma_P^2$
Between patients (total)	95	39505.99		

Solution continued on next page

As there is no replication (patients are only examined once), σ^2 cannot be estimated. (This term could perhaps be omitted from the model, but it should be retained as it is a part of the patient variation.)

$\sigma^2 + 4\sigma_P^2$ is estimated by 350.512.

Estimate of σ_D^2 is $\frac{1}{12}(464.349 - 350.512) = 9.49$.

Estimate of σ_C^2 would be $\frac{1}{48}(433.248 - 464.349)$, but σ_C^2 cannot be negative, so the estimate is taken as 0.

Thus we may conclude that the evidence suggests that there is no variability in the population of clinics, but there is variability in the population of doctors within clinics.

(iii) The research worker's $F_{1,94}$ result must have come from comparing the treatments with all other variation after it has been pooled. The sum of squares with 94 df would be 39505.99 – 4240.04 = 35265.95 with mean square 375.17. So his F statistic would be 4240.04/375.17 = 11.30, which is very close to the 0.1% critical point (11.57 for $F_{1,90}$).

Pooling in this way is not valid in general, as it ignores possible consistent sources of variation which are not part of the residual error.

(iv) We cannot estimate the basic measurement variation σ^2 (see part (ii) above).

The effects of treatments and clinics are confounded; treatments A and B should have been studied at all clinics (or, at the very least, both of them at one of the clinics).

To regard "clinics" as a random effect may be valid in a large area. "Doctors" seems less satisfactory as a random effect since at appears that only very small numbers of doctors were available at each clinic. We cannot say whether doctors choose their own patients genuinely at random; they may perhaps deliberately avoid – or deliberately choose – some whom they consider extreme for some reason. Presumably plenty of patients are available, but for the choice to be random it should have been made by someone else using a some form of anonymous list of those available.