EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)

GRADUATE DIPLOMA, 2001

Applied Statistics I

Time Allowed: Three Hours

Candidates should answer **FIVE** *questions.*

All questions carry equal marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use silent, cordless, non-programmable electronic calculators.

Where a calculator is used the ***method*** *of calculation should be stated in full.*

Note that $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$ *and that* ln *stands for* $\log_e$.

THERE IS A 14-PAGE APPENDIX FOR USE WITH THIS EXAMINATION PAPER

Each candidate must ensure (s)he has a copy of the Appendix.
The Appendix is to be handed in with the question paper and the answer book at the end of the examination. It must NOT be kept by the candidate.

This examination paper consists of 14 printed pages. This front cover is page 1. The reverse of the front cover is intentionally left blank and is page 2. Question 1 starts on page 3.

There are 8 questions altogether in the paper.

In addition, the Appendix is now included as pages 15-28 of the examination paper.
In the actual examination, it was supplied to candidates separately.

1. (a) A data analyst wants to try out an analysis of variance (ANOVA) using a regression package. He has two types of treatment, *A* and *B*, each at three levels. He creates the following dummy variables.

$$a_i = \begin{cases} 1 & \text{treatment } A \text{ at level } i \\ 0 & \text{treatment } A \text{ not at level } i \end{cases} \qquad i = 1, 2, 3$$

$$b_i = \begin{cases} 1 & \text{treatment } B \text{ at level } i \\ 0 & \text{treatment } B \text{ not at level } i \end{cases} \qquad i = 1, 2, 3$$

He tries fitting the model

$$y = \mu + \gamma_1 a_1 + \gamma_2 a_2 + \gamma_3 a_3 + \beta_1 b_1 + \beta_2 b_2 + \beta_3 b_3 + e \ .$$

His computer package tells him that it cannot fit this model.

(i) Explain why this model cannot be fitted.

(1)

(ii) How could the analyst adapt the model so as to allow the regression package to do a two-way analysis of variance?

(2)

(iii) What will the interpretation of the resulting parameter estimates be?

(5)

(iv) Assuming there are sufficient replicates in the data, how could he test to see if there is any interaction between *A* and *B*?

(4)

(b) Each of the following problems could occur in fitting a multiple linear regression model. For each of the problems suggest a possible solution. Justify your answers.

(i) The **X'X** matrix contains some very large values.

(ii) Multicollinearity exists.

(iii) The residuals are uncorrelated but have a non-constant variance.

(iv) The overall model is statistically significant, but none of the individual parameter estimates is significant.

(8)

2. A physician is interested in estimating the relationship between oxygen uptake and various physical characteristics of adolescent boys. A random sample of ten boys was selected and the following variables measured:

OXY	oxygen uptake
AGE	age in years
HEIG	height in centimetres
WEIG	weight in kilograms
CHES	chest circumference in centimetres.

Various models were fitted to the data, each of which included a constant term. For each model, the residual sum of squares is given (to 5 significant figures) in the **first table on the next page (page 5)**.

(i) Using the backward selection technique, determine which regressor variables best explain the variation in the response variable. Explain all your working and carefully justify your conclusions.

(12)

(ii) The **second table on the next page (page 5)** shows values of Cook's distance and leverage for the full model, i.e. the model including all possible regressor variables. Comment on these values. In the light of your findings, what other analyses would you carry out? Justify your answer.

(8)

Tables for question 2 are given on the next page

Terms in model (in addition to constant)	*Residual SS*
-	0.21296
AGE	0.20901
HEIG	0.020102
WEIG	0.12084
CHES	0.10309
AGE + HEIG	0.012758
AGE + WEIG	0.12079
AGE + CHES	0.10303
HEIG + WEIG	0.01516
HEIG + CHES	0.019428
WEIG + CHES	0.10250
AGE + HEIG + WEIG	0.0071493
AGE + HEIG + CHES	0.011133
AGE + WEIG + CHES	0.10249
HEIG + WEIG + CHES	0.014175
AGE + HEIG + WEIG + CHES	0.0069226

Observation	*Cook's distance*	*Leverage*
1	0.053428	0.3570
2	0.159995	0.5572
3	0.001054	0.2823
4	0.220663	0.3874
5	1.578700	0.7980
6	0.061630	0.3046
7	1.079330	0.5623
8	5.491473	0.9289
9	0.003968	0.4503
10	0.082316	0.3721

Turn over

3. Large software packages comprise a large number of modules, or subroutines. The modules undergo careful testing to remove any "bugs". Various specification parameters are recorded for a random selection of modules to assess how the numbers of faults vary with the specification parameters. The data in the table below record the numbers of detected faults, y, together with two global specification parameters, $SPEC1$ and $SPEC2$, for 12 software modules written by the same programmer.

y	$SPEC1$	$SPEC2$
3	14.154	0.1132
10	31.817	4.5437
4	2.203	5.1989
24	22.646	15.0614
5	8.585	2.6844
4	2.160	11.2151
43	53.517	22.5853
3	6.234	0.7164
2	2.858	0.8493
26	34.124	16.0000
3	2.484	5.6245
2	6.619	0.1385

A generalised linear model is fitted to the data using Poisson errors and a log link function.

(i) Explain why such a distribution and link function might be appropriate for these data.

(1)

(ii) A model is fitted with

$$\eta = \beta_0 + \beta_1 SPEC1 + \beta_2 SPEC2$$

The scaled deviance is 11.96. Comment on the apparent fit of the model to the data.

(2)

(iii) Figures 3.1 to 3.3 **on page 3 of the Appendix** show plots of the Pearson residual against the predicted value and each of the predictor values. Give detailed comments on the form of these plots.

(5)

(iv) Figures 3.4 and 3.5 **on page 4 of the Appendix** show a histogram and Normal plot of the Pearson residuals. Comment on these plots.

(4)

Question 3 is continued on the next page

(v) A second model was fitted with

$$\eta = \beta_0 + \beta_1 \log(SPEC1) + \beta_2 SPEC2$$

Comment on whether you think this was a sensible idea.

(1)

(vi) The scaled deviance from this second model is 4.3478. Comment on the apparent fit of this model.

(2)

(vii) Figures 3.6 to 3.10 **on pages 5 and 6 of the Appendix** show residual plots and a histogram and Normal plot of the Pearson residuals from this second model. Comment on the plots.

(2)

(viii) What further analyses would you carry out? Justify your answer.

(3)

4. The table below contains data on the effects of varying pressures on the density of cylindrical specimens made by dry pressing a ceramic compound. A mixture of aluminium oxide, polyvinyl alcohol and water was prepared, dried overnight and sieved. The resulting grains were pressed into cylinders at pressures from 2000 psi to 10000 psi and cylinder densities were calculated.

Pressure (psi)	*Density (g cm^{-3})*
2000	2.486
2000	2.479
2000	2.472
4000	2.558
4000	2.570
4000	2.580
6000	2.646
6000	2.657
6000	2.653
8000	2.724
8000	2.774
8000	2.808
10000	2.861
10000	2.879
10000	2.858

Details of a regression analysis are given **on pages 7 and 8 of the Appendix**. Making reference to this output, answer the following questions.

(i) Describe the apparent form of the relationship between the densities of the cylindrical specimens and the pressures used in their manufacture.

(2)

(ii) Give a suitable estimate of the percentage of the variability in the densities of the specimens explained by a linear term in the pressure.

Interpret the value of this estimate in the context of the experiment.

Briefly discuss the advantages and disadvantages of using R^2 and adjusted R^2 when assessing results from a regression analysis.

(3)

(iii) Is the following statement correct? Give a detailed justification of your answer.

> "The value 0.00004867, the estimate of the gradient of the true regression line, is very small. This means that there is little effect of pressure on the density of the cylindrical specimens."

(3)

Question 4 is continued on the next page

(iv) Use the residual plots to check appropriate assumptions of the model that has been fitted. Describe carefully how you are using the plots in checking each assumption. State any assumptions you have been unable to check.

(7)

(v) Given the nature of the design of this study, the assumption of linearity can be checked more efficiently.

(a) Explain what aspect of the experimental design enables this check to be used.

(b) Explain how the partitioning of the error sum of squares (0.00515) from the analysis of variance table given in the output enables this test to be carried out.

(5)

5. The data in the table below give the measurements of the inorganic phosphorus, organic phosphorus and estimated plant-available phosphorus in samples of soils at 18 degrees Celsius.

Inorganic phosphorus (units not supplied) *INORGP*	*Organic phosphorus (units not supplied)* *ORGP*	*Plant-available phosphorus (p.p.m)* *PLANTP*
0.4	53	64
0.4	23	60
3.1	19	71
0.6	34	61
4.7	24	54
1.7	65	77
9.4	44	81
10.1	31	93
11.6	29	93
12.6	58	51
10.9	37	76
23.1	46	96
23.1	50	77
21.6	44	93
23.1	56	95
1.9	36	54
26.8	58	168
29.9	51	99

Output from a regression analysis is given **on pages 9 to 11 of the Appendix**. The variable names are *INORGP*, *ORGP* and *PLANTP* respectively. Making reference to this output, answer the following questions.

(i) Describe the apparent relationships between the three variables, treating *PLANTP* as the response.

(4)

(ii) Describe how the model comparison statistics AR^2 and Mallows' C_p can best be used to compare the fit of different linear models [**Note:** these are shown on the output as "Adj. R-sq" and "C-p" respectively].

Use the values of these two model comparison statistics to choose a linear model to describe the relationship between the response and the two potential predictors (*INORGP* and *ORGP*). Explain your reasoning.

(5)

Question 5 is continued on the next page

(iii) For the model relating *PLANTP* to *INORGP* and fitted to the full dataset:

(a) Write down the theoretical model used (for the *i*th observation), defining all parameters and variables.

(b) perform a formal test of significance for the parameter corresponding to the variable *INORGP*, giving a full specification of your null and alternative hypotheses.

(5)

(iv) Give a detailed explanation of the observed effect of removing observation 17. What does the output indicate about observation 17?

(6)

6. (a) Consider the second-order moving average process $\{X_t\}$ where

$$X_t = Z_t + 0.7Z_{t-1} - 0.2Z_{t-2}$$

and $\{Z_t\}$ is a white noise process with $E(Z_t) = 0$ and $\text{Var}(Z_t) = \sigma_z^2$.

Obtain the mean, variance and autocorrelation function of $\{X_t\}$ and hence show that the process is second-order stationary.

(12)

(b) Describe the form of the correlogram for the following general types of series.

(i) A random series.

(ii) An alternating series.

(iii) A series with increasing trend.

(iv) A series exhibiting seasonal fluctuations.

(8)

7. The following table gives data from a replicated two-factor experiment.

A1		*A2*		*A3*	
B1	*B2*	*B1*	*B2*	*B1*	*B2*
7	4	12	5	9	4
6	9	7	7	13	8
8	7	9	5	10	6
10	9	10	8	12	4
7	7	12	7	14	4
38	36	50	32	58	26

(a) Consider both factors to be fixed.

(i) Write down the form of the model.

(3)

(ii) Complete a suitable analysis of variance.

(7)

(iii) Fully describe your conclusions.

(6)

(b) State the distinction between a *fixed* and a *random* factor. Briefly explain how your analysis would have changed if *B* had been a random factor. (There is no need to carry out any calculations.)

(4)

8. (a) Compare the aims of principal component analysis and cluster analysis.

(4)

(b) Data are available from the 1996 Olympic Games, describing the complete results of all 31 competitors in the decathlon event. The variables are defined as follows.

m100	100 metres, time in seconds
longj	long jump, in metres
shot	shot put, in metres
hjump	high jump, in metres
m400	400 metres, in seconds
m110h	110 metres hurdles, in seconds
discus	discus, in metres
polevlt	pole vault, in metres
jav	javelin, in metres
m1500	1500 metres, in seconds

The output **on pages 12 to 14 of the Appendix** gives the results of multivariate analyses of these data. Making reference to this output, answer the following questions.

(i) Describe the correlations between the variables, identifying any possible clusters of variables.

(2)

(ii) Using the results of the principal components analysis, draw a scree plot. State how many principal components you would use to summarise the data, justifying your answer.

(4)

(iii) Interpret the first four principal components.

(4)

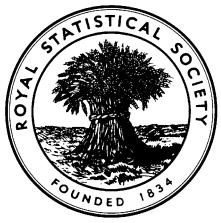
(iv) A cluster analysis was performed on the 31 observations, with dissimilarities defined as the raw Euclidean distances between the points in the dataset described above. Comment on the validity of such a cluster analysis.

(2)

(v) Compare and contrast the information given in the dendrogram and the labelled plots of the first four principal components.

(4)

EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)



GRADUATE DIPLOMA, 2001

Applied Statistics I

APPENDIX

Each candidate must have a copy of this Appendix.

The Appendix consists of FOURTEEN pages.

This front cover is page 1.

The reverse of the front cover, which is intentionally left blank, is page 2.

The text of the Appendix starts on page 3.

The Appendix is to be handed in with the candidate's examination paper and answer book at the end of the examination. It is NOT to be removed by the candidate.

Output for question 3. This output is printed on this page and the next three pages

Residual plots from model with spec1 and spec2

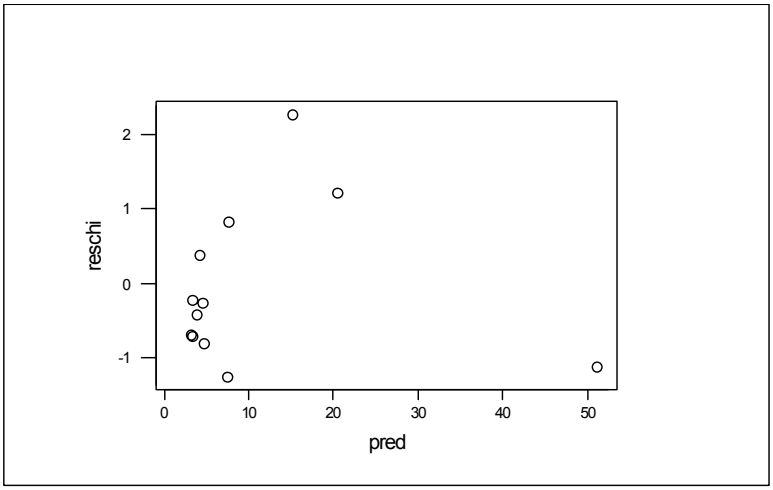


Figure 3.1. Residuals against predicted values

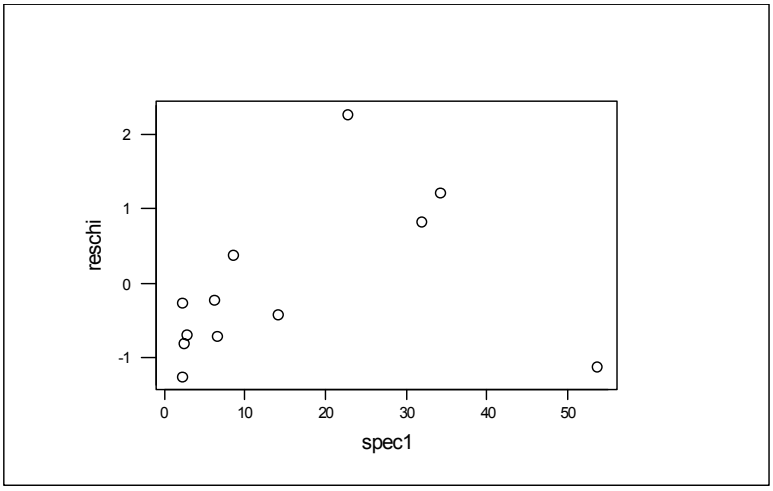


Figure 3.2. Residuals against SPEC1

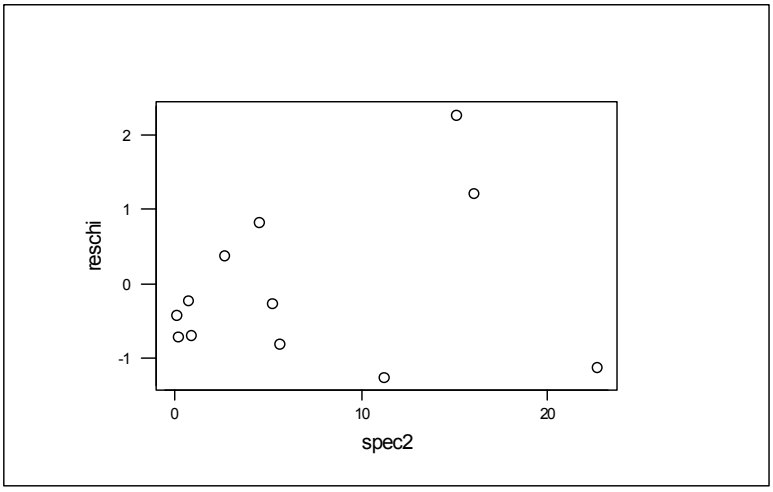


Figure 3.3. Residuals against SPEC2

Output for question 3 (contd)

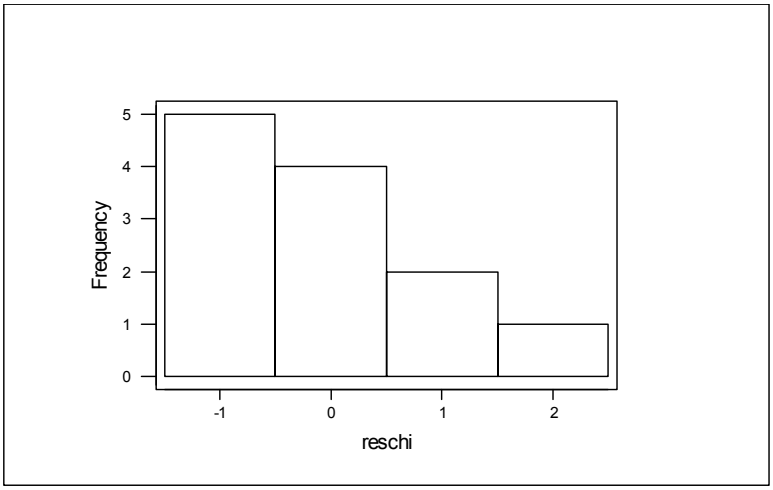


Figure 3.4. Histogram of residuals

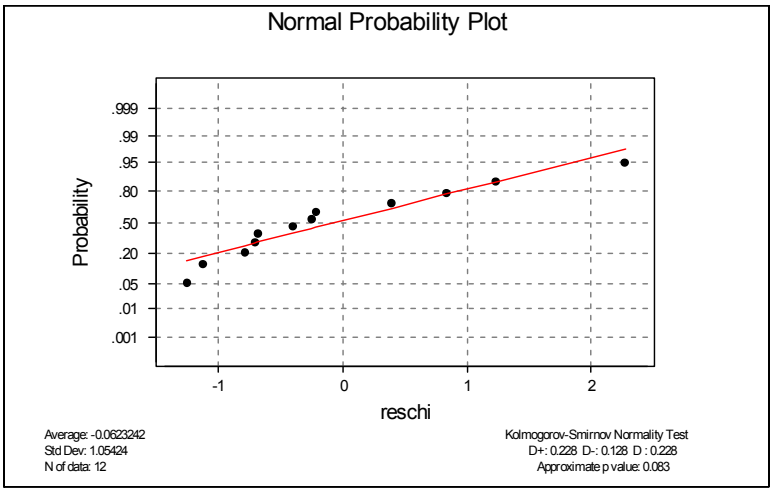


Figure 3.5. Normal plot of residuals

Output for question 3 (contd)

Residual plots from model with log(spec1) and spec2

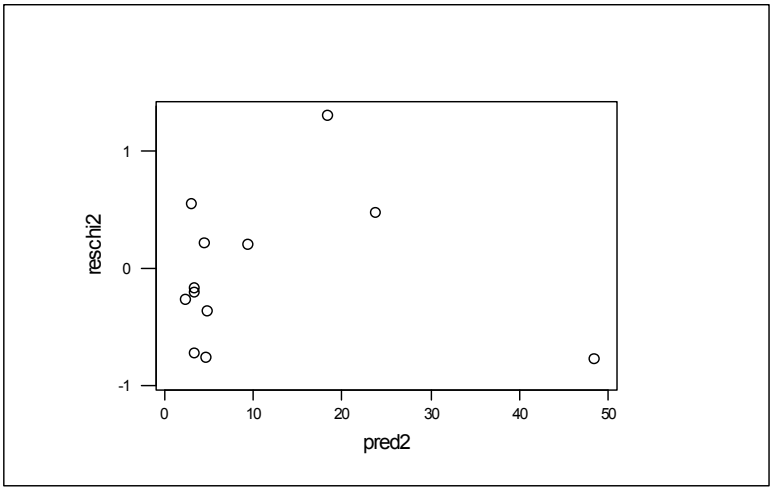


Figure 3.6. Residuals against predicted values

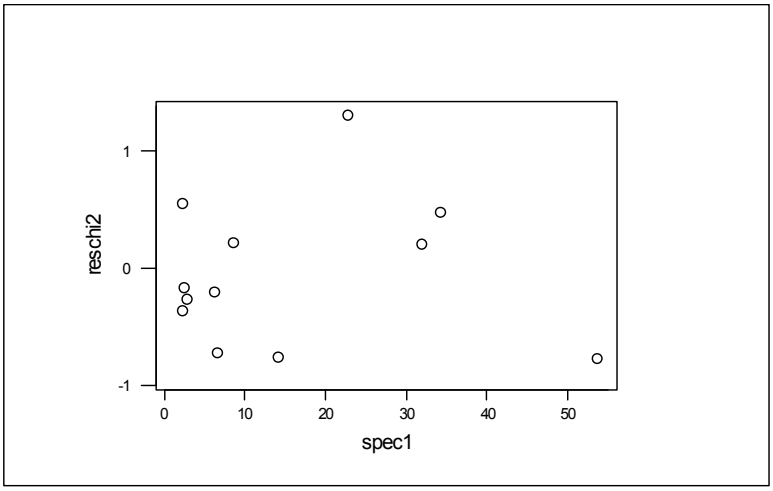


Figure 3.7. Residuals against SPEC1

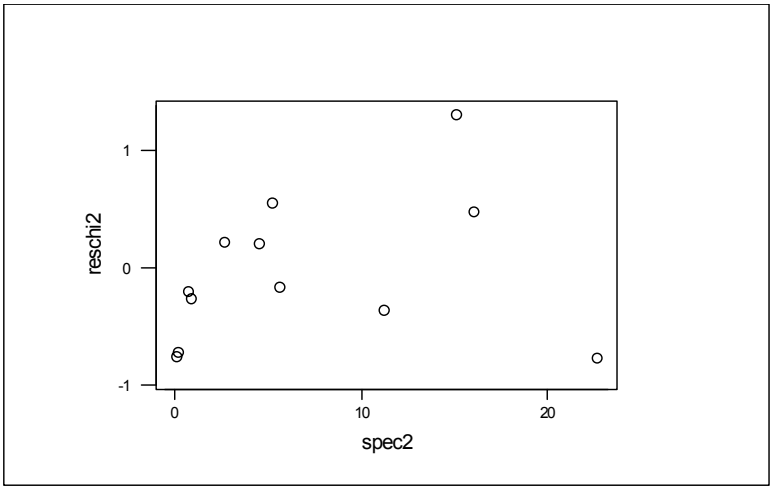


Figure 3.8. Residuals against SPEC2

Output for question 3 (contd)

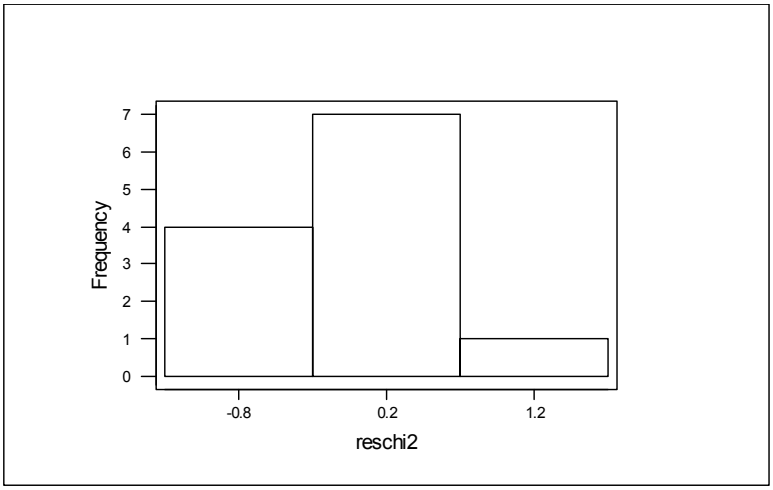


Figure 3.9. Histogram of residuals

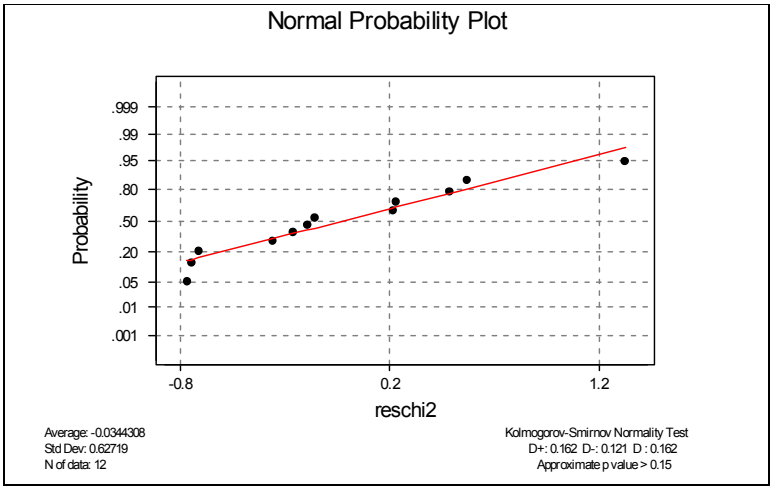


Figure 3.10. Normal plot of residuals

Output for question 4. This output is printed on this page and the next page

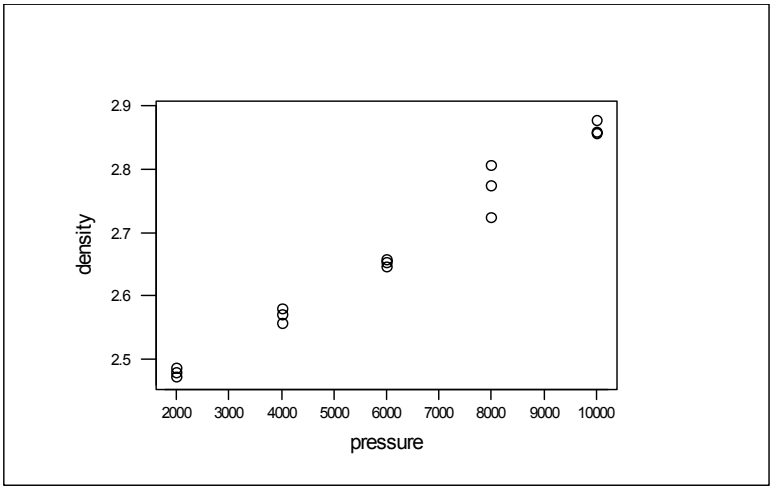


Figure 4.1. Scatter plot of density against pressure

```
Regression Analysis


The regression equation is
density = 2.37 +0.000049 pressure

Predictor       Coef       Stdev    t-ratio        p
Constant     2.37500     0.01206     197.01    0.000
pressure  0.00004867  0.00000182      26.78    0.000

s = 0.01991     R-sq = 98.2%     R-sq(adj) = 98.1%

Analysis of Variance

SOURCE       DF          SS          MS          F        p
Regression    1     0.28421     0.28421    717.06    0.000
Error        13     0.00515     0.00040
Total        14     0.28937

Unusual Observations
Obs. pressure    density        Fit  Stdev.Fit   Residual    St.Resid
 10      8000    2.72400    2.76433    0.00630   -0.04033      -2.14R
 12      8000    2.80800    2.76433    0.00630    0.04367       2.31R

R denotes an obs. with a large st. resid.
```

Output for question 4 (contd)

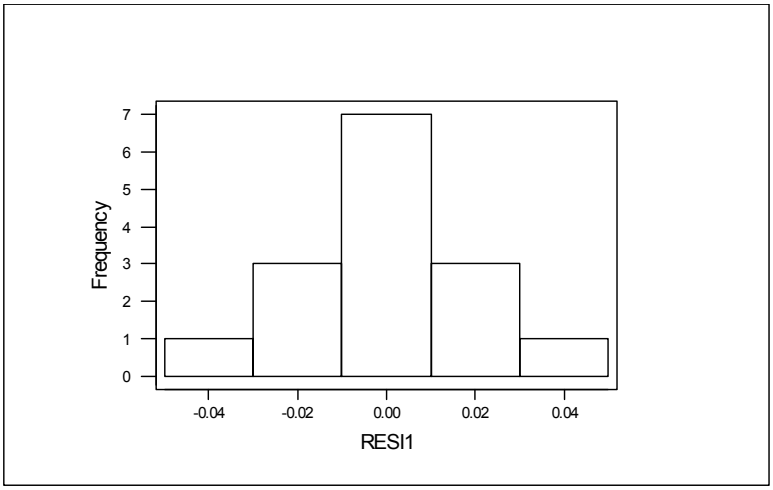


Figure 4.2. Histogram of residuals

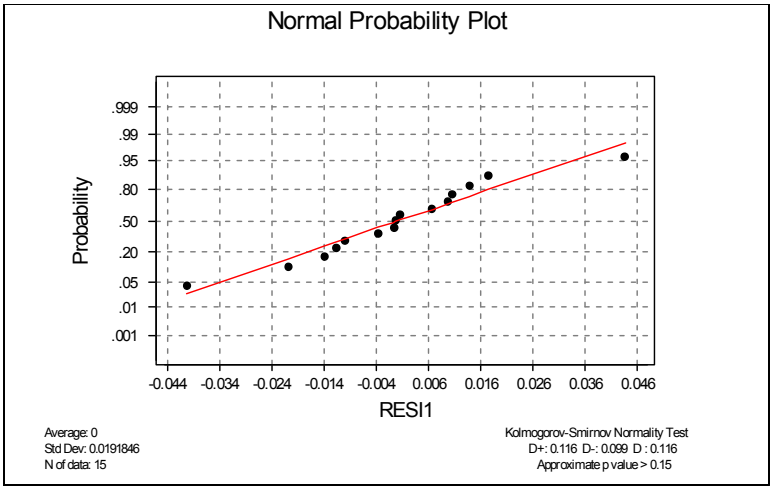


Figure 4.3. Normal plot of residuals

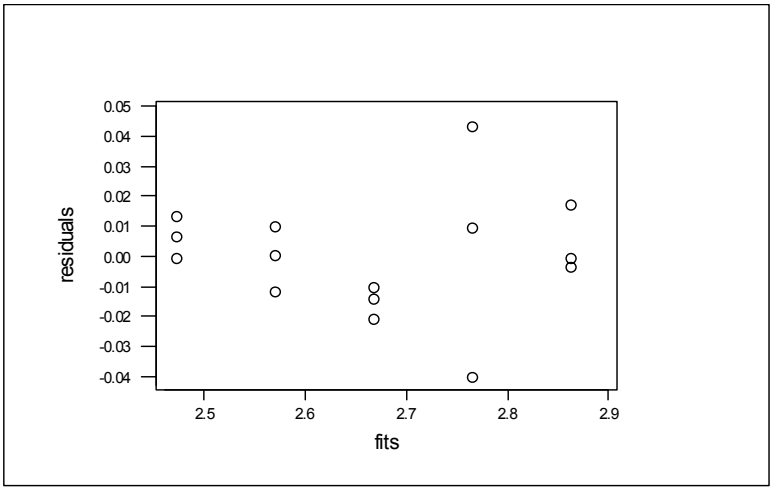


Figure 4.4. Plot of residuals against fitted values

Output for question 5. This output is printed on this page and the next two pages

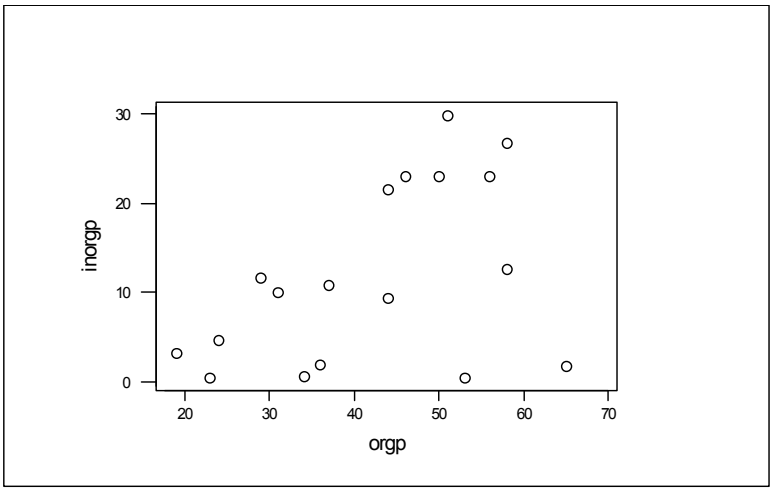


Figure 5.1. Scatter plot of INORGP against ORGP

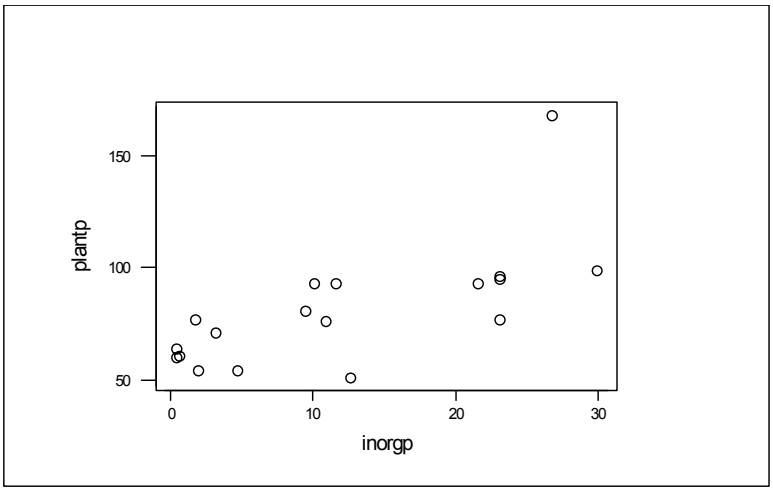


Figure 5.2. Scatter plot of PLANTP against INORGP

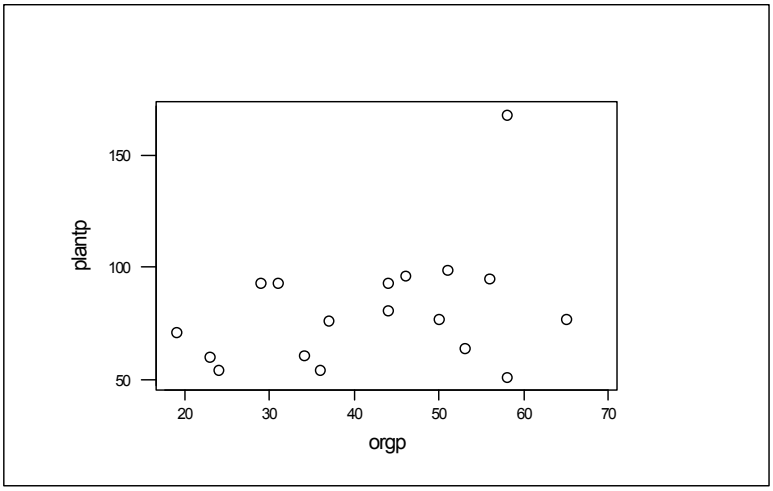


Figure 5.3. Scatter plot of PLANTP against ORGP

Output for question 5 (contd)

Best Subsets Regression

Response is plantp

Vars	R-sq	Adj. R-sq	C-p	s	inorgp	orgp
1	48.1	44.8	1.0	20.051	X	
1	12.6	7.1	11.3	26.020		X
2	48.2	41.3	3.0	20.678	X	X

Regression Analysis (including all observations)

The regression equation is
plantp = 59.3 + 1.84 inorgp

Predictor	Coef	Stdev	t-ratio	p
Constant	59.259	7.420	7.99	0.000
inorgp	1.8434	0.4789	3.85	0.001

s = 20.05 R-sq = 48.1% R-sq(adj) = 44.8%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	5957.0	5957.0	14.82	0.001
Error	16	6432.6	402.0		
Total	17	12389.6			

Unusual Observations

Obs.	inorgp	plantp	Fit	Stdev.Fit	Residual	St.Resid
17	26.8	168.00	108.66	8.54	59.34	3.27R

R denotes an obs. with a large st. resid.

Output for question 5 (contd)

Regression Analysis (excluding observation 17)

```
Regression Analysis


The regression equation is
plantp = 62.6 + 1.23 inorgp

Predictor       Coef       Stdev    t-ratio         p
Constant      62.569       4.452      14.05     0.000
inorgp        1.2291      0.3058       4.02     0.001

s = 11.92       R-sq = 51.9%     R-sq(adj) = 48.6%

Analysis of Variance

SOURCE       DF          SS          MS          F         p
Regression    1      2295.2      2295.2      16.15     0.001
Error        15      2131.2       142.1
Total        16      4426.5

Unusual Observations
Obs.   inorgp     plantp        Fit  Stdev.Fit   Residual    St.Resid
 10      12.6      51.00      78.06       2.93     -27.06      -2.34R

R denotes an obs. with a large st. resid.
```

Output for question 8. This output is printed on this page and the next two pages

Correlations (Pearson)

	m100	longj	shot	hjump	m400	m110h	discus	polevlt	jav
longj	-0.405								
shot	-0.196	0.251							
hjump	0.101	0.285	0.117						
m400	0.685	-0.215	-0.087	0.210					
m110h	0.530	-0.325	-0.155	-0.036	0.351				
discus	-0.280	0.306	0.349	0.216	-0.004	-0.380			
polevlt	-0.337	0.328	0.069	0.043	-0.224	-0.061	0.161		
jav	-0.344	0.281	0.163	0.083	-0.027	-0.429	0.297	0.185	
m1500	-0.225	-0.002	-0.214	0.207	-0.098	-0.205	0.307	-0.008	0.013

Output from principal component analysis

Principal Component Analysis

Eigenanalysis of the Correlation Matrix

Eigenvalue	3.0084	1.5649	1.2961	1.0357	0.8810	0.7503
Proportion	0.301	0.156	0.130	0.104	0.088	0.075
Cumulative	0.301	0.457	0.587	0.691	0.779	0.854

Eigenvalue	0.5154	0.4176	0.3195	0.2112
Proportion	0.052	0.042	0.032	0.021
Cumulative	0.905	0.947	0.979	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
m100	0.472	0.322	-0.065	-0.018	-0.074	-0.027
longj	-0.375	0.169	-0.217	0.326	0.008	0.390
shot	-0.222	0.220	-0.504	-0.288	0.515	-0.160
hjump	-0.088	0.586	0.121	0.345	0.135	0.437
m400	0.315	0.538	-0.043	-0.086	-0.309	-0.183
m110h	0.416	0.039	-0.171	0.371	0.136	-0.251
discus	-0.341	0.378	0.121	-0.184	0.176	-0.507
polevlt	-0.254	-0.083	-0.216	0.653	-0.237	-0.489
jav	-0.323	0.157	-0.101	-0.275	-0.700	0.006
m1500	-0.149	0.105	0.760	0.103	0.141	-0.191

Variable	PC7	PC8	PC9	PC10
m100	0.149	-0.207	0.087	-0.768
longj	0.586	0.371	0.162	-0.132
shot	-0.322	0.097	0.393	-0.083
hjump	-0.428	-0.237	-0.240	0.098
m400	0.211	0.019	0.370	0.539
m110h	-0.178	0.665	-0.321	0.023
discus	0.338	-0.068	-0.533	-0.033
polevlt	-0.112	-0.328	0.192	-0.060
jav	-0.359	0.328	-0.103	-0.220
m1500	-0.128	0.303	0.427	-0.179

Output for question 8 (contd)

Labelled plots of principal component scores for the 31 athletes

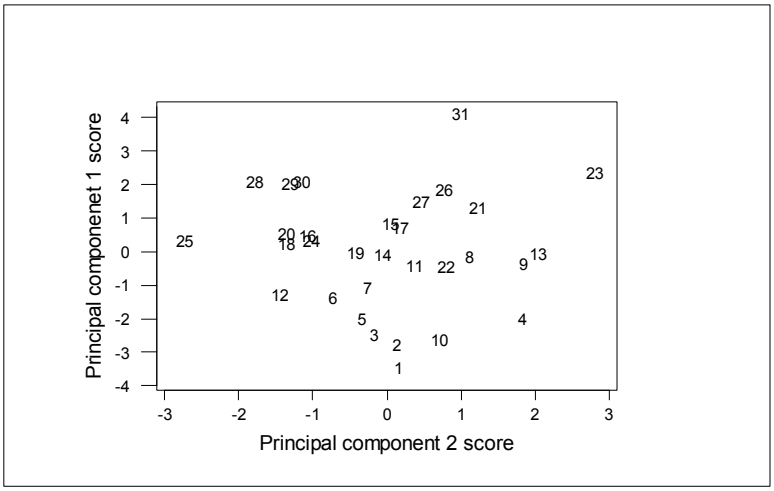


Figure 8.1. Plot of principal component score 1 against principal component score 2

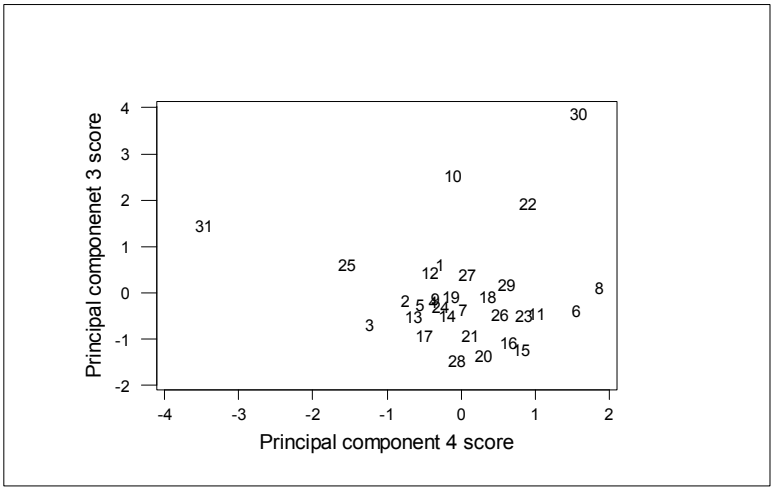


Figure 8.2. Plot of principal component score 3 against principal component score 4

Output for question 8 (contd)

Output from cluster analysis of observations, using Euclidean distance and average linkage

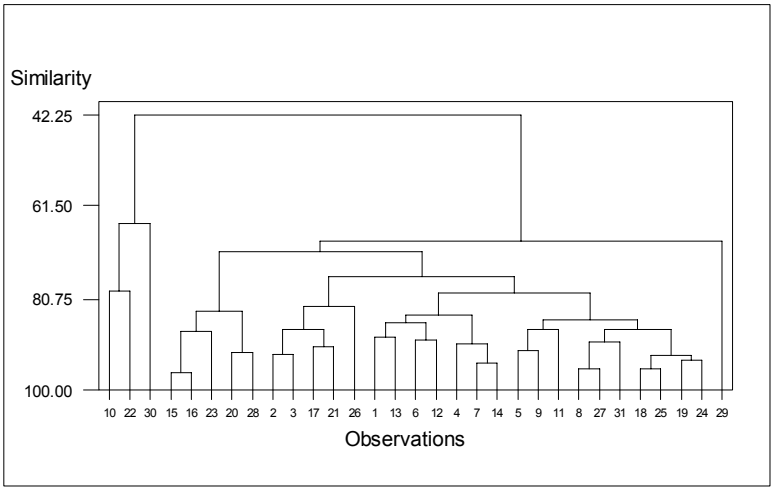


Figure 8.3. Dendrogram from cluster analysis