

EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY



GRADUATE DIPLOMA, 2009

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 calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

The notation \log denotes logarithm to base e .

Logarithms to any other base are explicitly identified, e.g. \log_{10} .

Note also that $\binom{n}{r}$ is the same as nC_r .

This examination paper consists of 13 printed pages, **each printed on one side only**.

This front cover is page 1.

Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. (i) Define the terms *stationarity* and *weak stationarity* in the context of time series analysis. Explain why weak stationarity is often used in practice. (2)
- (ii) Write down a formula for the k th sample autocorrelation for a time series. Explain how the k th partial autocorrelation differs from the k th autocorrelation. Explain also how the autocorrelation function (ACF) and partial ACF (PACF) may be used in identifying an appropriate model for time series data. (3)
- (iii) (a) Define an autoregressive series of order p , i.e. $AR(p)$.
 (b) Define a moving average series of order q , i.e. $MA(q)$.
 (c) Sketch a typical ACF and PACF for $AR(1)$ and $MA(1)$ series. (4)
- (iv) The computer output below describes the sample ACF and PACF from analyses of four time series labelled TS1, TS2, TS3 and TS4. Each of these has 50 observed values. They are not necessarily stationary time series. By considering the ACF and PACF suggest, for each time series, a possible model that could be used as a basis for further analysis. Justify your answer. (11)

LAG	TS1		TS2		TS3		TS4	
	AC	PAC	AC	PAC	AC	PAC	AC	PAC
1	-0.14	-0.14	0.93	0.94	0.18	0.19	-0.57	-0.57
2	-0.06	-0.08	0.86	-0.09	-0.61	-0.68	0.64	0.48
3	0.01	-0.01	0.79	-0.21	-0.63	-0.67	-0.59	-0.26
4	-0.13	-0.14	0.75	0.15	0.14	-0.51	0.56	0.14
5	0.09	0.05	0.70	-0.02	0.87	0.79	-0.41	0.40
6	0.11	0.13	0.64	-0.25	0.18	-0.13	-0.38	-0.10
7	-0.26	-0.24	0.54	-0.23	-0.53	-0.08	-0.38	-0.10
8	0.10	0.01	0.46	0.07	-0.57	-0.24	0.30	0.10
9	0.14	0.18	0.39	-0.13	0.10	-0.19	-0.30	-0.13
10	0.02	0.10	0.33	-0.20	0.75	-0.20	0.21	-0.16
11	0.13	0.11	0.26	-0.00	0.13	-0.57	-0.26	-0.20
12	-0.02	0.07	0.17	-0.18	-0.43	-0.32	0.12	-0.22
13	-0.23	-0.27	0.09	-0.08	-0.48	-0.35	-0.15	-0.05
14	-0.00	-0.15	0.01	-0.03	0.10	-0.33	0.04	-0.17
15	-0.05	-0.08	-0.11	-0.41	0.63	0.04	-0.17	-0.48
16	0.05	0.04	-0.05	-0.03	0.09	0.12	0.00	-0.32
17	-0.07	-0.32	-0.20	-0.50	-0.37	-0.28	-0.06	-0.13
18	0.09	0.36	-0.27	0.14	-0.41	-0.35	0.04	-0.33
19	-0.10	-0.36	-0.33	-0.50	0.13	0.17	0.04	0.10
20	0.04	0.10	-0.37	-0.42	0.54	0.22	-0.03	-0.20
21	0.05	-0.09	-0.41	-0.55	0.05	-0.33	0.13	-0.13
22	-0.10	-0.22	-0.45	-0.78	-0.33	-0.46	-0.13	-0.87
23	-0.11	-0.27	-0.46	-0.20	-0.34	0.28	0.23	0.67

2. In many instances of linear modelling, a response variable y might be dependent on more than one predictor variable. Thus a set of variables x_i ($i = 1, 2, \dots, p$) could be used to predict y through the general linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

where the β_i are model parameters and ε is an error term.

- (i) State clearly all the assumptions made in fitting such a model. (2)
- (ii) Write down the equivalent matrix formulation of the model, and state the form of the least squares estimators for the parameters in the model. (3)
- (iii) These least squares estimators have some very useful properties.
- (a) State the properties they possess irrespective of the distribution of the errors.
- (b) State the extra properties they possess if the errors are independent and Normally distributed. (3)
- (iv) Explain why highly dependent predictor variables can cause problems in fitting such a model. What methods can be used to try to overcome such problems? (3)
- (v) Explain why an adjusted R^2 value is often preferred to R^2 when comparing models. (2)
- (vi) Explain what is meant by *influential observations* and why they can be a problem. Describe some of the diagnostics that can be obtained from statistical packages to detect influential observations. (7)

3. An economist is studying salaries for employees in a large company. He has data on 434 employees. The variables are as follows.

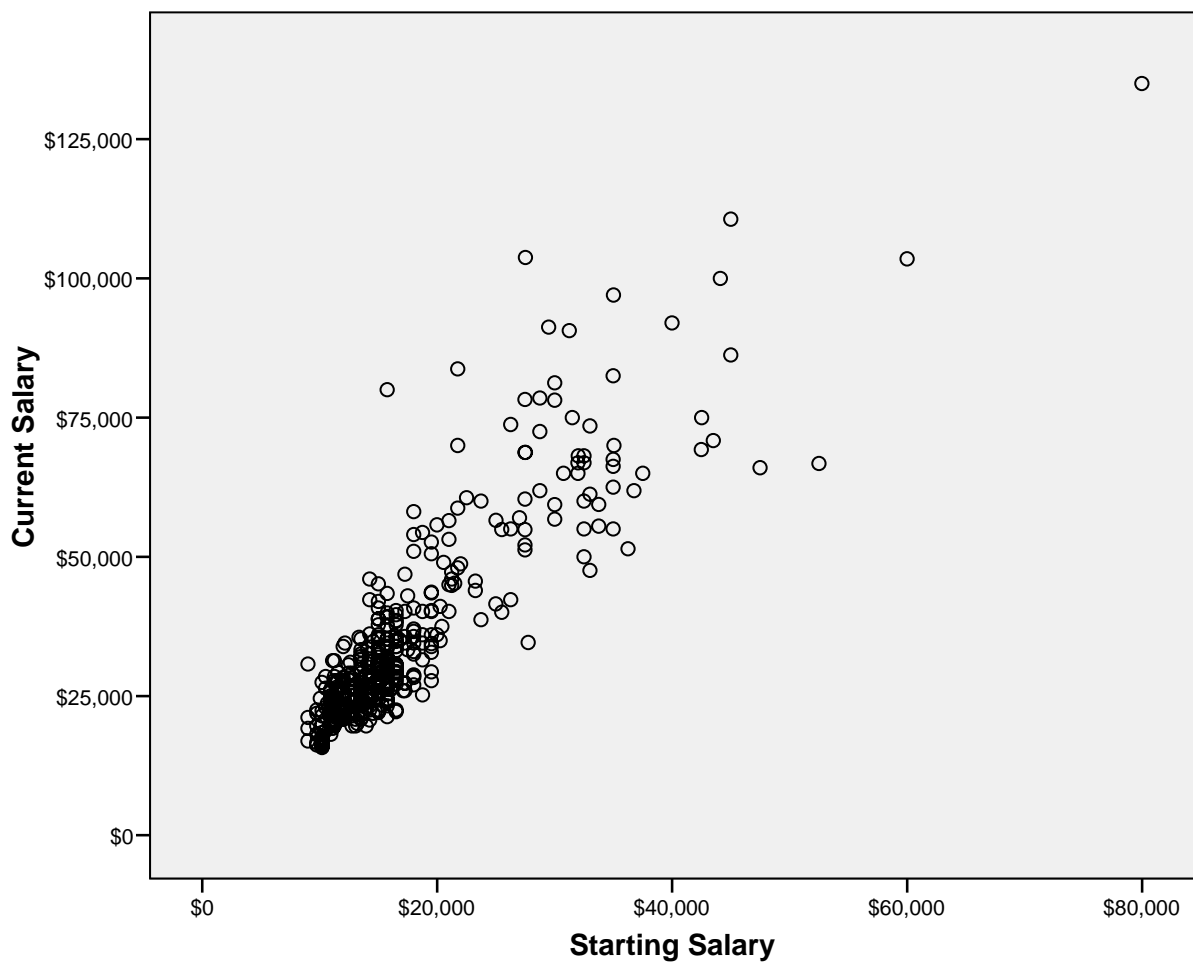
Current salary in dollars

Starting salary, when the employee joined the company, in dollars

Sex of employee

Grade of employee (clerk or team leader or manager)

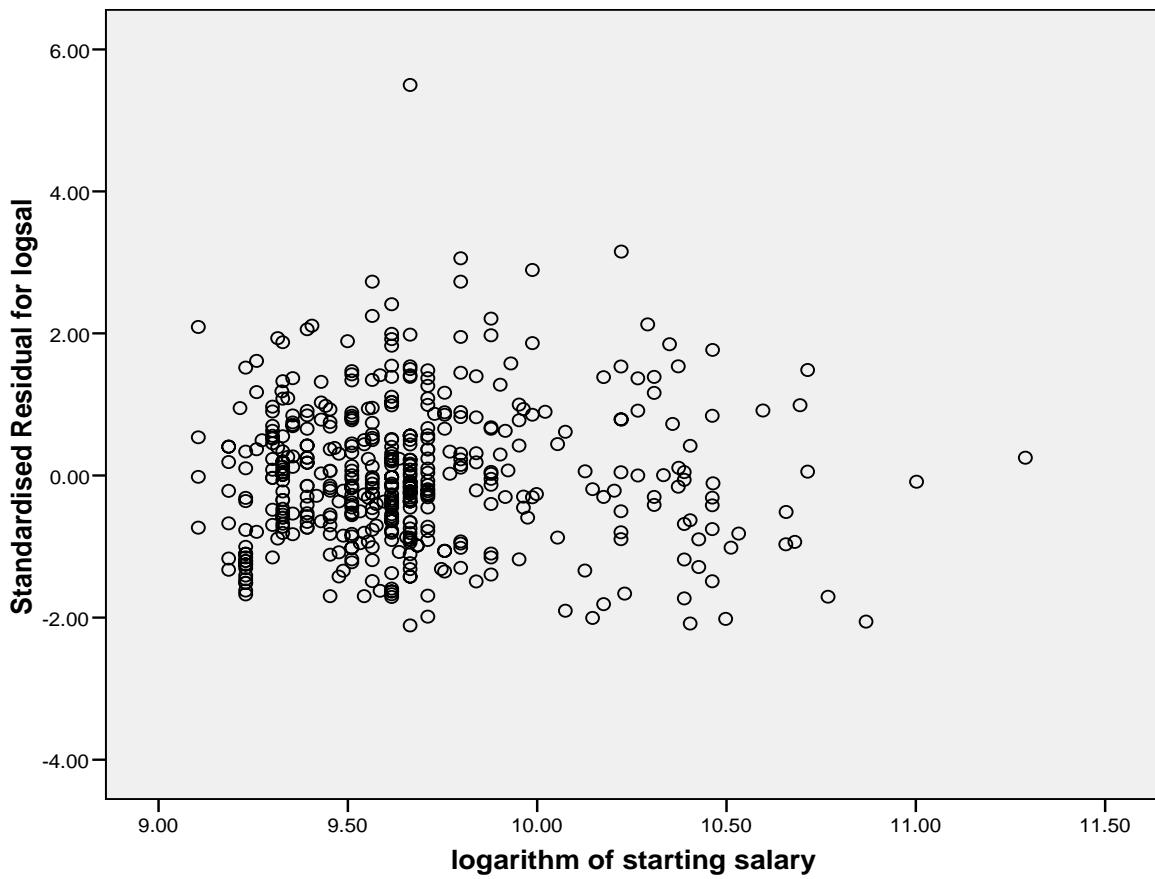
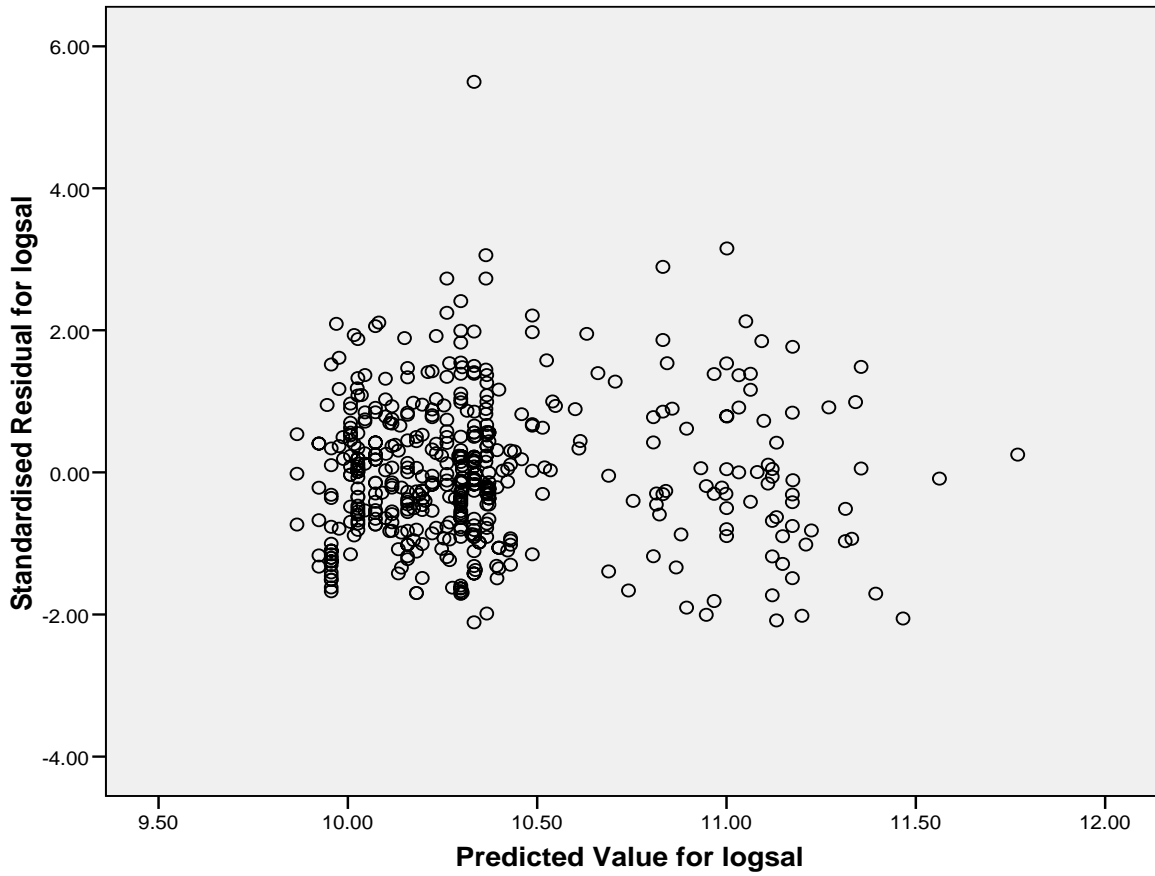
The economist wants to construct a model to predict current salary from the other variables, and is convinced that current salary is related to starting salary. He shows the following scatter plot as evidence of this relation.



Question 3 is continued on the next two pages

- (i) Someone tells the economist that it would be better if he scaled all the salary values by dividing by 1,000 to make the numbers more manageable. He is worried that this might affect the analysis, changing the values and statistical significance of the parameter estimates obtained. What would your advice be and why? (2)
- (ii) Someone else advises the economist to model the logarithm of the current salary value rather than the raw value. Why might this be a good idea? (2)
- (iii) The economist suggests that if he is using the logarithm of the current salary then maybe he should use the logarithm of the starting salary. What would you advise? (3)
- (iv) The economist has a computer program that will do multiple regression, but he does not know how to model factors using this program. Write down the form of a suitable multiple regression model to include all three predictor variables, and using the logarithms of the salary values. Explain each term in the model and how it relates to the original variables. (4)
- (v) The economist obtains a multiple regression model using the logarithms of both of the variables describing salaries. The response variable is called logsal. He produces a plot of the standardised residuals against the predicted values obtained from the model, and a plot of the standardised residuals against the logarithm of starting salary. Interpret these plots in a way that he could understand. **The plots are shown on the next page.** (6)
- (vi) Someone else says that they have found an interaction between job category and logarithm of starting salary.
- (a) Show how this could be included in your multiple regression model.
- (b) Explain what such an interaction would mean. (3)

Question 3 is continued on the next page, which shows the plots for part (v)



4. (i) Briefly discuss the relative merits of forward selection and backward elimination as applied to model selection in multiple linear regression. (5)
- (ii) Three process variables X_1 , X_2 and X_3 can be adjusted in a chemical plant, and each variable might affect the yield Y from the plant. Twenty-one observations have been made on Y at different values of X_1 , X_2 and X_3 . It is required to find the best combination of X_1 , X_2 and X_3 for prediction of the yield, Y . The table below shows residual sums of squares from various linear models fitted to the data.

Variables in model	Residual sum of squares
–	2069.24
X_1	319.12
X_2	483.15
X_3	1738.44
X_1, X_2	188.80
X_1, X_3	309.14
X_2, X_3	475.06
X_1, X_2, X_3	178.83

- (a) Use forward selection to choose a model. Show your working. (3)
- (b) Use backward elimination to choose a model. Show your working. (3)
- (c) Which model would you recommend and why? (2)
- (d) What other information would you like to have in order to suggest a "good" model? (3)
- (e) How would you check that your final chosen model was a good fit to the data? (4)

5. (a) Explain the purpose of cluster analysis and discuss briefly the decisions that need to be made when selecting a method of cluster analysis. (5)

- (b) A researcher has collected data relating to 68 inner-city pregnant women. The following variables were measured.

mstress	mother's stress level
fstress	father's report of the mother's stress level
fsupp1	early level of support by father
fsupp2	later level of support by father
mdep1	early level of mother's depression
mdep2	later level of mother's depression
age	mother's age (years)
ethnic	mother's ethnicity
mstat	mother's marital status
parity	parity – the number of children the mother has

A higher score denotes a higher value of stress, support or depression.

Stress was measured near the end of the pregnancy. Support and depression were measured twice: in the first stage of pregnancy (labelled "early") and then again in the later stages of pregnancy (labelled "later").

With the exception of the variable *fstress*, all data were obtained from the mothers.

The correlation matrix is given below.

	mstress	fstress	fsupp1	fsupp2	mdep1	mdep2	age	ethnic	mstat	parity
mstress	1.00	0.17	-0.28	-0.27	0.23	0.50	0.06	-0.19	-0.18	0.19
fstress	0.17	1.00	-0.18	-0.18	0.10	0.14	0.16	-0.09	0.01	0.13
fsupp1	-0.28	-0.18	1.00	0.44	-0.34	-0.42	0.04	-0.16	0.12	-0.11
fsupp2	-0.27	-0.18	0.44	1.00	-0.17	-0.41	-0.24	-0.14	0.24	-0.17
mdep1	0.23	0.10	-0.34	-0.17	1.00	0.55	-0.35	0.11	-0.04	0.10
mdep2	0.50	0.14	-0.42	-0.41	0.55	1.00	-0.09	0.13	-0.20	0.16
age	0.06	0.16	0.04	-0.24	-0.35	-0.09	1.00	-0.02	0.05	0.26
ethnic	-0.19	-0.09	-0.16	-0.14	0.11	0.13	-0.02	1.00	-0.34	0.31
mstat	-0.18	0.01	0.12	0.24	-0.04	-0.20	0.05	-0.34	1.00	-0.12
parity	0.19	0.13	-0.11	-0.17	0.10	0.16	0.26	0.31	-0.12	1.00

Question 5 is continued on the next page

- (i) Briefly outline the main purpose of principal component analysis for such a set of data. (2)
- (ii) State the conditions under which it would be appropriate to include the two variables "ethnic" and "mstat" in a principal component analysis. Justify your answer. (2)
- (iii) Would you recommend carrying out a principal component analysis on the covariance matrix? Justify your answer. (2)
- (iv) Discuss the main correlations in the eight variables excluding ethnicity and marital status. (3)
- (v) The results of a correlation-based principal component analysis on the eight variables in (iv) are presented below, but only seven eigenvalues have been printed out. Deduce the missing eigenvalue and interpret the results. (6)

Eigenvalues

1.53 0.90 0.87 0.76 0.56 0.43 0.33

Eigenvectors

	1	2	3	4	5	6	7	8
mstress	-0.40	0.06	-0.24	-0.01	0.75	0.31	0.12	0.32
fstress	-0.22	0.27	0.36	0.86	0.02	-0.04	0.11	-0.07
fsuppl	0.43	0.05	-0.36	0.18	0.32	-0.68	0.30	0.05
fsupp2	0.40	-0.26	-0.35	0.40	0.16	0.34	-0.56	-0.20
mdep1	-0.39	-0.43	-0.17	0.19	-0.26	-0.36	-0.32	0.55
mdep2	-0.51	-0.17	-0.12	-0.08	0.18	-0.33	-0.17	-0.73
age	-0.01	0.71	0.03	-0.16	0.08	-0.24	-0.63	0.14
parity	-0.21	0.37	-0.72	0.11	-0.45	0.18	0.22	-0.08

6. (i) A random variable X has mean μ and variance σ^2 , and $f(X)$ is a function of X . Show that, to a first approximation, $f(X)$ has mean $f(\mu)$ and variance $\sigma^2 (f'(\mu))^2$ where $f'(\mu) = \left. \frac{df(x)}{dx} \right|_{x=\mu}$. What is a necessary condition for the approximations to be reasonable? (3)

(ii) Use the result in (i) to show that, if X has a Poisson distribution with parameter λ , then \sqrt{X} has approximately a mean of $\sqrt{\lambda}$ and a variance of $1/4$. (2)

(iii) In an attempt to model the flow of traffic, an analyst counts the number of vehicles passing a given point on a road in a 10-minute interval on 5 occasions on each of 3 different days. The results are as given below.

	<i>Number of vehicles passing in a 10 minute interval</i>				
<i>Day 1</i>	5	5	4	1	3
<i>Day 2</i>	14	9	5	12	10
<i>Day 3</i>	5	4	4	3	8

The analyst performs a one-way analysis of variance after applying the square root transformation to these data. Part of the analysis is given below.

<i>Source</i>	<i>Sums of Squares</i>
Between days	4.440
Within days	3.046
Total	7.486

(a) Write down the statistical model that is being used for this analysis. (4)

(b) Complete the analysis of variance table and state your conclusions. (6)

(c) Confirm that the within days variance of the transformed data is approximately $1/4$. (2)

(d) An alternative method of analysis would be to use a generalised linear model. Write down this model and state whether you would expect the results from this model to differ from your answer in (b). (3)

7. The data in the table below show the lengths of time (in seconds) taken by rats to find their way out of a maze. Three different breeds of rats were chosen and the rats were kept under one of two conditions for three months before being placed in the maze. In one condition they were highly restricted in the way they could move around, whereas in the second condition they could move quite freely. Four different rats were used for each possible combination of breed and condition.

		<i>Breed A</i>	<i>Breed B</i>	<i>Breed C</i>
Condition	<i>Free</i>	24 16	107 101	130 110
		39 33	81 98	107 102
	<i>Restricted</i>	125 93	121 132	95 108
		127 89	156 138	98 134

For information: $\sum_i \sum_j \sum_k x_{ijk} = 2364$, $\sum_i \sum_j \sum_k x_{ijk}^2 = 264128$, where x_{ijk} is the time for the k th rat from the j th breed in condition i .

- (i) Explain why 'condition' is a fixed factor. (2)
- (ii) Describe two scenarios, one where 'breed' would be a fixed factor and one where it would be a random factor. (3)
- (iii) Write down the two different two-way models (including interactions) for each of the two scenarios in (ii). Explain all the terms in each model. (6)
- (iv) Obtain the analysis of variance assuming 'breed' to be a fixed factor. Include graphical representations if appropriate. (6)
- (v) Describe how the nature of possible conclusions would have been different if 'breed' had been a random factor. (3)

8. A doctor is investigating the effect of a woman's age on the success of an IVF (in vitro fertilisation) procedure. She has randomly selected 10 women aged under 35 and 10 women aged at least 35. From hospital records she has obtained the following data, which record the numbers of eggs obtained from the women and the numbers that were fertilised during one IVF procedure. She wants to investigate the effect of the woman's age on the probability of an egg being successfully fertilised. She calls this probability the "fertilisation rate".

Women aged under 35		Women aged at least 35	
<i>Number of eggs</i>	<i>Number of fertilised eggs</i>	<i>Number of eggs</i>	<i>Number of fertilised eggs</i>
10	9	7	6
9	7	10	7
7	5	9	5
5	3	8	4
10	9	6	4
7	7	5	1
9	5	7	4
8	8	6	4
7	2	5	2
7	5	7	5

- (i) Carry out a suitable exploratory analysis to see whether the fertilisation rate might depend on the woman's age. (4)
- (ii) Let n_i denote the number of eggs and x_i the number of fertilised eggs for the i th woman. Let π_i denote the fertilisation rate for the i th woman.
- (a) Explain why a binomial distribution may be valid to model the data. (2)
- (b) Write down the expression for the log likelihood of the observed data, assuming a binomial distribution with different fertilisation rates for each woman. Identify the logit function in your expression. (2)

Question 8 is continued on the next page

- (iii) The data are analysed using a generalised linear model, with the logit link. The model assumes constant fertilisation rate within each age group, so contains a constant and age as a covariate. Age is coded as 1 for older women, and 0 for younger women. Part of the output from a computer program is given below.

Deviance = 28.26	(1/df) Scaled Deviance = 1.57
Variance function: $V(u) = u*(1-u/eggs)$	[Binomial]
Link function : $g(u) = \log(u/(eggs-u))$	[Logit]

- (a) Explain why the highlighted value 1.57 is useful, and how it is derived from the other numerical value in the output. (2)
- (b) Explain what the highlighted expressions $V(u)$ and $g(u)$ are and how their formulae are obtained. (2)
- (c) The estimated value of the coefficient for age in the generalised linear model is -0.744 and the estimate of the constant is 1.150.
Obtain estimates of the predicted success rates for the two types of women. (4)
- (d) For the model which contains only the constant (i.e. does not take age into account), the value for the scaled deviance is 32.65. State, with reasoning, whether the effect of woman's age is statistically significant. (2)
- (e) Someone else has modelled these data but coded younger women as age = 1 and older women as age = 0. Explain how the results and estimates would be different from those given above. (2)