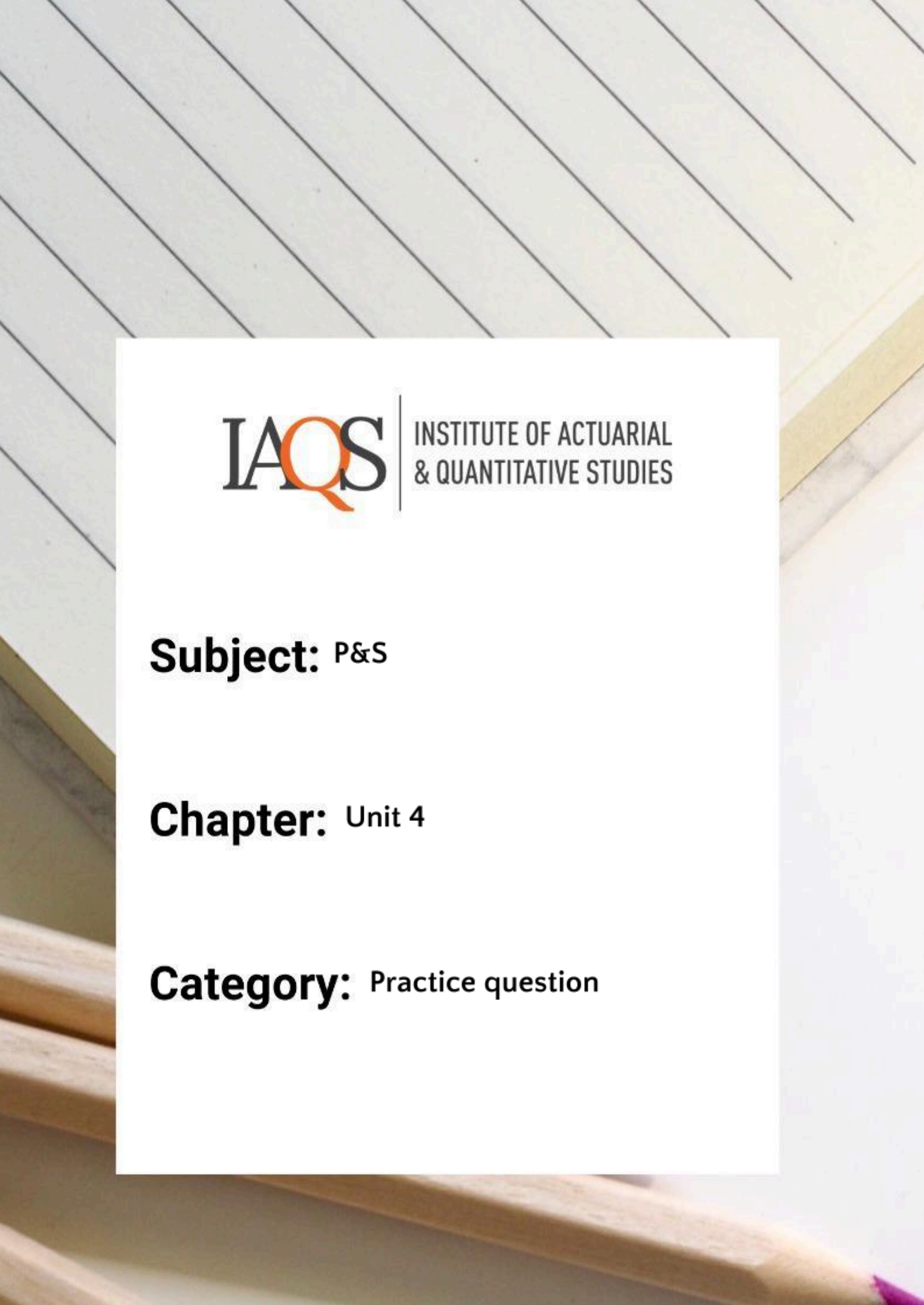


IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: P&S

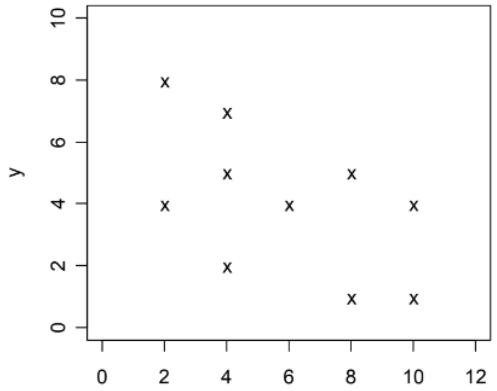
Chapter: Unit 4

Category: Practice question

Correlation analysis and regression

1. CT6 October 2011 Question 10

Consider a situation in which integer-valued responses (y) are recorded at ten values of an integer-valued explanatory variable (x). The data are presented in the following scatter plot:



For these data:

$$\sum x = 58, \sum x^2 = 420, \sum y = 41, \sum y^2 = 217, \sum xy = 202$$

(i) (a) Calculate the value of the coefficient of determination (R^2) for the data.
(b) Determine the equation of the fitted least-squares line of regression of y on x .

(ii) Calculate a 95% confidence interval for the slope of the underlying line of regression of y on x.

(iii) (a) Calculate an estimate of the expected response in the case x = 9 .
(b) Calculate the standard error of this estimate.

Suppose the observation (x = 10, y = 8) is added to the existing data. The coefficient of determination is now $R^2 = 0.07$.
(iv) Comment briefly on the effect of the new observation on the fit of the linear model.

2. CT6 April 2012 Question 13

The quality of primary schools in eight regions in the UK is measured by an index ranging from 1 (very poor) to 10 (excellent). In addition the value of a house price index for these eight regions is observed.
The results are given in the following table:

Unit 4

PRACTICE QUESTION

Region i	1	2	3	4	5	6	7	8	Sum
School quality index x_i	7	8	5	8	4	9	6	9	56
House price index y_i	195	195	170	190	150	190	200	210	1500

The last column contains the sum of all eight columns.

From these values we obtain the following results:

$$\sum x_i y_i = 10,695; \quad \sum x_i^2 = 416; \quad \sum y_i^2 = 283,750$$

(i) Calculate the correlation coefficient between the index of school quality and the house price index.

You can assume that the joint distribution of the two random variables is a bivariate normal distribution.

(ii) Perform a statistical test for the null hypothesis that the true correlation coefficient between the school quality index and the house price index is equal to 0.8 against the alternative that the correlation coefficient is smaller than 0.8, by calculating an approximate p -value.

(iii) Fit a linear regression model to the data, by considering the school quality index as the explanatory variable. You should write down the model and estimate all parameters.

(iv) Calculate the coefficient of determination 2 R for the regression model obtained in part (iii).

(v) Provide a brief interpretation of the slope of the regression model obtained in part (iii).

3. CT6 October 2012 Question 13

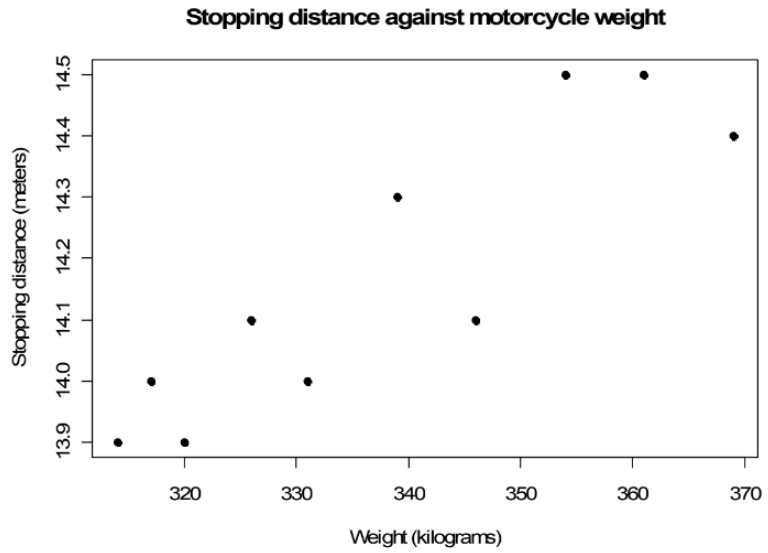
The following data give the weight, in kilograms, of a random sample of 10 different models of similar motorcycles and the distance, in metres, required to stop from a speed of 20 miles per hour.

Weight x	314	317	320	326	331	339	346	354	361	369
Distance y	13.9	14.0	13.9	14.1	14.0	14.3	14.1	14.5	14.5	14.4

For these data: $\sum x = 3{,}377$, $\sum x^2 = 1{,}143{,}757$, $\sum y = 141.7$,
$\sum y^2 = 2{,}008.39$, $\sum xy = 47{,}888.6$

Also: $S_{xx} = 3{,}344.1$, $S_{yy} = 0.501$, $S_{xy} = 36.51$

A scatter plot of the data is shown below.



(i) (a) Comment briefly on the association between weight and stopping distance, based on the scatter plot.

(b) Calculate the correlation coefficient between the two variables.

(ii) Investigate the hypothesis that there is positive correlation between the weight of the motorcycle and the stopping distance, using Fisher's transformation of the correlation coefficient. You should state clearly the hypotheses of your test and any assumption that you need to make for the test to be valid.

(iii) (a) Fit a linear regression model to these data with stopping distance being the response variable and weight the explanatory variable.

(b) Calculate the coefficient of determination for this model and give its interpretation.

(c) Calculate the expected change in stopping distance for every additional 10 kilograms of motorcycle weight according to the model fitted in part (iii)(a).

Unit 4

PRACTICE QUESTION

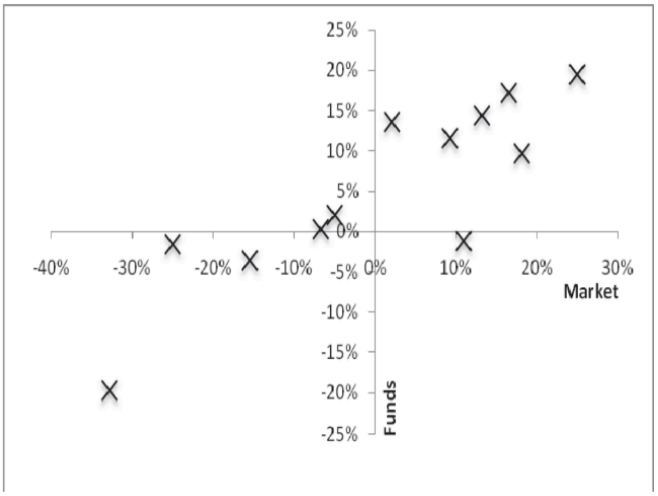
3. CT6 October 2013 Question 10

An analyst wishes to compare the results from investing in a certain category of hedge funds, f, with those from the stock market, x. She uses an appropriate index for each, which over 12 years each produced the following returns (in percentages to one decimal place).

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Market (x)	–5.0	–15.4	–25.0	16.6	9.2	18.1	13.2	2.0	–32.8	25.0	10.9	–6.7
Funds (f)	2.1	–3.7	–1.6	17.3	11.6	9.7	14.4	13.7	–19.8	19.5	–1.2	0.3

$$\sum x = 0.101,\ \sum x^2 = 0.3612,\ \sum f = 0.622,\ \sum f^2 = 0.1710,\ \sum xf = 0.1989$$

It is assumed that observations from different years are independent of each other.
Below is a scatter plot of market returns against fund returns for each year.



(i) Comment on the relationship between the two series.

The hedge fund industry often claims that hedge funds have low correlation with the stock market.

(ii) (a) Calculate the correlation coefficient between the two series.

(b) Test whether the correlation coefficient is significantly different from 0.

(iii) Calculate the parameters for a linear regression of the fund index on the market index.

(iv) Calculate a 95% confidence interval for the underlying slope coefficient for the linear model in part (iii).

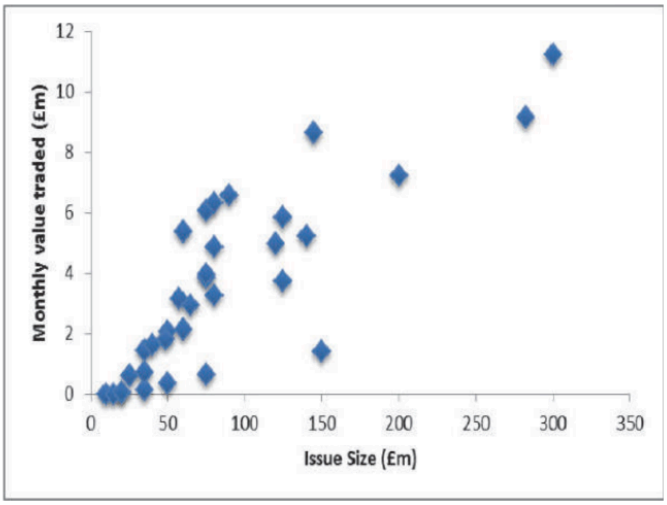
(v) Comment on your answers to parts (ii)(b) and (iv).

Unit 4

PRACTICE QUESTION

4. CT6 April 2014 Question 10

An analyst is instructed to investigate the relationship between the size of a bond issue and its trading volumes (value traded). The data for 33 bonds are plotted in the following chart.



(i) Comment on the relationship between issue size and value traded.

The analyst denotes issue size by s and monthly value traded by v. He calculates the following from the data:

$$\sum s_i = 2,843.7,\ \sum s_i^2 = 397,499.8,\ \sum v_i = 115.34,\ \sum v_i^2 = 689.37,\ \sum s_i v_i = 15,417.75$$

(ii) (a) Determine the correlation coefficient between s and v .

(b) Perform a statistical test to determine if the correlation coefficient is significantly different from 0.

(iii) Determine the parameters of a linear regression of v on s and state the fitted model equation.

(iv) State the outcome of a statistical test to determine whether the slope parameter in part (iii) differs significantly from zero, justifying your answer.

A colleague suggests that the central part of the data, with issue sizes between £50m and £150m, seem to have a greater spread of value traded and without the bonds in the upper and lower tails the linear relationship would be much weaker.

(v) Comment on the colleague's observation.

Unit 4

PRACTICE QUESTION

5. CT6 September 2014 Question 10

An insurer has collected data on average alcohol consumption (units per week) and cigarette smoking (average number of cigarettes per day) in eight regions in the UK.

Region, I	1	2	3	4	5	6	7	8	Average
Alcohol units per week, xi	15	25	21	29	13	18	21	17	19.875
Cigarettes per day, yi	4	8	8	10	6	9	7	5	7.125

For these observations we obtain:

$$\sum x_i y_i = 1{,}190; \quad \sum x_i^2 = 3{,}355; \quad \sum y_i^2 = 435$$

(i) Calculate the coefficient of correlation between alcohol consumption and cigarette smoking.

(ii) Calculate a 95% confidence interval for the true correlation coefficient. You may assume that the joint distribution of the two random variables is a bivariate normal distribution.

(iii) Fit a linear regression model to the data, by considering alcohol consumption as the explanatory variable. You should write down the model and estimate the values of the intercept and slope parameters.

(iv) Calculate the coefficient of determination R^2 for the regression model in part (iii).

(v) Give an interpretation of R^2 calculated in part (iv).

6. CT6 October 2015 Question 5

An insurance company is accused of delaying payments for large claims. To investigate this accusation a sample of 25 claims is considered. In each case the claim size x_i (in £) and the time y_i (in days) taken to pay the claim are recorded.

Assume that the claim size and the time taken to pay the claim are normally distributed. In the sample the following statistics have been observed:

Unit 4

PRACTICE QUESTION

$$\sum_{i=1}^{25}(x_i - \bar{x})^2 = 5{,}116{,}701 \qquad \sum_{i=1}^{25}(y_i - \bar{y})^2 = 61.44$$

$$\sum_{i=1}^{25}(x_i - \bar{x})(y_i - \bar{y}) = 2{,}606.96$$

(i) Calculate the correlation coefficient between the claim sizes, x_i , and the times taken to pay the claim, y_i.

(ii) Perform a statistical test of the hypothesis that the correlation between claim size and time until payment is zero against the alternative that the correlation is different from zero.

7. CT6 October 2015 Question 11

A property agent carries out a study on the relationship between the age of a building and the maintenance costs, X, per square meter per annum based on a sample of 86 buildings. In the sample denote by x_i the annual maintenance costs per square meter for building i .

In a first step the sample is divided into new and old buildings. The maintenance costs are summarised in the following table:

	sample size n	$\sum x_i$	$\sum x_i^2$
new buildings	25	100	800
old buildings	61	300	2200

(i) Perform a test for the null hypothesis that the variance of the maintenance costs of new buildings is equal to the variance of the maintenance costs for old buildings, against the alternative that the variance of the maintenance costs of new buildings is larger. Use a significance level of 5%.

(ii) Perform a test of the null hypothesis that the mean of the maintenance costs of new buildings is equal to the mean of the maintenance costs for old buildings, against the alternative of different means. Use a significance level of 5%.

To obtain further insight into the relationship between age and maintenance costs for old buildings the agent wishes to carry out a linear regression analysis. Let A denote the age of a building and X denote the annual maintenance costs per square metre. The agent uses the model $E[X] = \gamma A + \beta$.

The agent has the following summary data for the age a_i and costs x_i of the 61 old buildings in the sample.

$$\sum_{i=1}^{61} a_i = 4{,}500, \quad \sum_{i=1}^{61} a_i x_i = 30{,}000 \text{ and } \quad \sum_{i=1}^{61} a_i^2 = 506{,}400$$

(iii) Estimate the correlation coefficient $\rho(A, X)$ between age A and maintenance costs X.

(iv) Estimate the parameters γ and β.

8. CT6 April 2016 Question 11

A car magazine published an article exploring the relationship between the mileage (in units of 1,000 miles) and the selling price (in units of £1,000) of used cars. The following data were collected on 10 four-year-old cars of the same make.

Car	1	2	3	4	5	6	7	8	9	10
Mileage, x	42	29	51	46	38	59	18	32	22	39
Price, y	5.3	6.1	4.7	4.5	5.5	5.0	6.9	5.7	5.8	5.9

$$\sum x = 376, \sum x^2 = 15600, \sum y = 55.4, \sum y^2 = 311.44, \sum xy = 2014.5$$

(i) (a) Determine the correlation coefficient between x and y.
(b) Comment on its value.

A linear model of the form $y = a + \beta x + \varepsilon$ is fitted to the data , where the error terms (ε) independently follow a $N(0, \sigma^2)$ distribution, with σ^2 s being an unknown parameter.

(ii) Determine the fitted line of the regression model.

(iii) (a) Determine a 95% confidence interval for β

The article suggests that there is a 'clear relationship' between mileage and selling price of the car.

(b) Comment on this suggestion based on the confidence interval obtained in part (iii)(a).

(iv) Calculate the estimated difference in the selling prices for cars that differ in mileage by 5,000 miles.

9. CT6 April 2017 Question 10

A geologist is trying to determine what causes sand granules to have different sizes. She measures the gradient of nine different beaches in degrees, g, and the diameter in mm of the granules of sand on each beach, d.

g	0.63	0.70	0.82	0.88	1.15	1.50	4.40	7.30	11.30
d	0.17	0.19	0.22	0.235	0.235	0.30	0.35	0.42	0.85

$$\sum g = 28.68,\ \sum g^2 = 206.2462,\ \sum d = 2.97,\ \sum d^2 = 1.33525,\ \sum gd = 15.55855$$

(i) Determine the linear regression equation of d on g.

The geologist assumes that the error terms in the linear regression are normally distributed.

(ii) Perform a test to determine whether the slope coefficient is significantly different from zero.

(iii) Determine a 95% confidence interval for the mean estimate of d on a beach with a slope of exactly 3 degrees.

(iv) (a) Plot the data from the table above.

(b) Comment on the plot suggesting what the geologist might do to improve her analysis.

10. CT6 September 2017 Question 10

A company leases animals, which have been trained to perform certain tasks, for use in the movie industry. The table below gives the number of tasks that each of nine monkeys in a random sample can perform, along with the number of years the monkeys have been working with the company.

Name	Hellion	Freeway	SuSu	Henri	Jo	Peepers	Cleo	Jeep	Maggie
Years	10	8	6.5	6	5	1.5	0.5	0.5	0.4
Tasks	28	24	28	28	27	23	15	6	23

The random variable Y_i denotes the number of years and T_i the number of tasks for each monkey i = 1,…,9.

$$\sum y_i = 38.4,\ \sum y_i^2 = 270.16,\ \sum y_i t_i = 1011.2,\ \sum t_i = 202,\ \sum t_i^2 = 4976$$

(i) Explain the roles of response and explanatory variables in a linear regression.

(ii) Determine the correlation coefficient between Y and T.

Unit 4

PRACTICE QUESTION

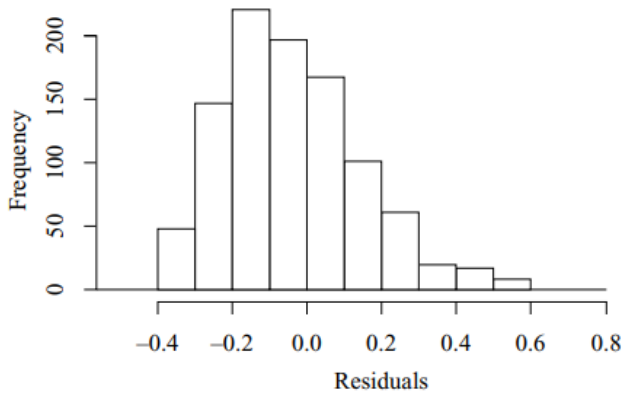
(iii) Perform a statistical test using Fisher's transformation to determine whether the population correlation coefficient is significantly different from zero.

(iv) Determine the parameters of a linear regression, including writing down the equation.

11. CS1A September 2019 Question 6

An actuary is asked to check a linear regression calculation performed by a trainee. The trainee reports a least squares slope parameter estimate of $\hat{b} = 13.7$ and a sample correlation coefficient r = –0.89.

(i) Justify why this suggests that the trainee has made an error. In a different simple linear regression model, a histogram of the residuals is shown below



(ii) Comment on the validity of the assumptions of the linear model

x	0	1	2	3	4	5	6	7	8	9
y	−1.35	−4.96	− 9.20	−13.15	−16.70	−21.23	−25.14	−28.44	−33.68	−37.39

for which

$$\bar{y} = -19.124, \quad \sum_{i=1}^{10}(y_i - \bar{y})^2 = 1{,}329.523, \quad \sum_{i=1}^{10}(x_i - \bar{x})^2 = 82.5,$$

$$\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = -331.05$$

Unit 4

PRACTICE QUESTION

A linear model of the form $y = \alpha + \beta x + e$ is fitted to the data, where the error terms (e) independently follow a $N(0, \sigma^2)$ distribution, and where a, b and s2 are unknown parameters.
(iii) Determine the fitted line of the regression model.

(iv) Calculate a 95% confidence interval for the predicted mean response if x = 11.

(v) Comment on the width of a 95% confidence interval for the predicted mean response if x = 3.5, as compared to the width of the interval in part (iv), without calculating the new interval.

12. CS1A September 2020 Question 9

For an empirical investigation into the amount of rent paid by tenants in a town, data on income X and rent Y have been collected. Data for a total of 300 tenants of one-bedroom flats have been recorded. Assume that X and Y are both Normally distributed with expectations μ_X and μ_Y, and variances σ_X^2 and σ_Y^2. S_X and S_Y are the sample standard deviation for random samples of X and Y, respectively.

The random variable Z_X is defined as

$$Z_X = 299\frac{S_X^2}{\sigma_X^2}.$$

(i) State the distribution of Z_X and all of its parameters.

(ii) Write down the expectation and variance of Z_X.

(iii) Explain why the distribution of Z_X is approximately Normal.

(iv) Calculate values of an approximate 2.5% quantile and 97.5% quantile of the distribution of Z_X using your answers to parts (ii) and (iii).

In the collected sample, the mean income is \$1,838 with a realised sample standard deviation of \$211, the mean rent is \$608 with a realised sample standard deviation of \$275 and $\Sigma x_i y_i = 348 \times 10^6$

(v) Calculate a 95% confidence interval for the mean income.

(vi) Calculate a 95% confidence interval for the mean rent.

Unit 4

PRACTICE QUESTION

(vii) Calculate an approximate 95% confidence interval for the variance of income using your answer to part (iv).

(viii) Identify which one of the following options gives the correct form of the equation for the simple linear regression model of rent on income, including any assumptions required for statistical inference.

A1 $y_i = a + bx_i$

A2 $y_i = a + bx_i + z_i \, with \, E[z_i] = 0$

A3 $y_i = a + bx_i + z_i \, with \, z_i \sim \chi^2, \, 299 \, df$

A4 $y_i = a + bx_i + z_i \, with \, z_i \sim N(0, \, \sigma 2)$

(ix) Calculate estimates of the slope and the intercept of the model in part (viii) based on the above data for the 300 tenants.

13. CS1A September 2020 Question 5

Consider a regression model in which the response variable Yi is linked to the explanatory variable Xi by the following equation:

$$Y_i = a + bX_i + e_i \, , \, i = 1, \ldots, n$$

assuming that the error terms ei are independent and Normally distributed with expectation 0 and variance σ^2. In a sample of size n = 10, the following statistics have been observed:

$$\sum_{i=1}^{n} x_i = 141, \quad \sum_{i=1}^{n} y_i = 127,$$

$$\sum_{i=1}^{n} x_i^2 = 2{,}014, \quad \sum_{i=1}^{n} y_i^2 = 1{,}629, \quad \sum_{i=1}^{n} x_i y_i = 1{,}810.$$

(i) Calculate values for S_{xx}, S_{yy}, $and \, S_{xy}$.

(ii) Write down, using your answers to part (i), the value of Pearson's correlation coefficient between the variables $X_i \, and \, Y_i$

(iii) Calculate estimates of the parameters a and b in the regression model.

Unit 4

PRACTICE QUESTION

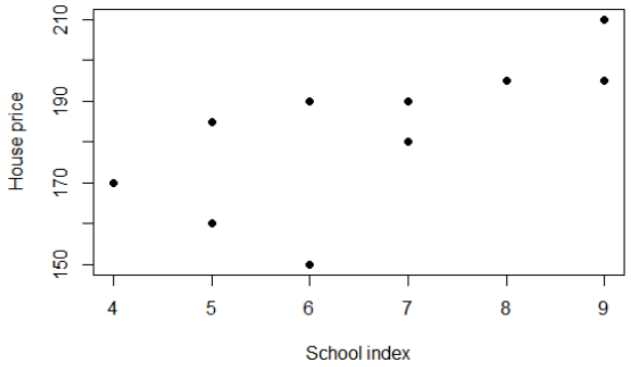
14. CS1A September 2020 Question 10

It is thought that house prices in certain areas are correlated with the quality of schools in the same areas. A study has been carried out in ten regions where average house prices and school quality indices ranging from 1 (very poor) to 10 (excellent) have been recorded:

Region i	1	2	3	4	5	6	7	8	9	10
School index x_i	9	5	7	6	4	9	7	8	5	6
House prices y_i (£1,000s)	210	185	190	190	170	195	180	195	160	150

$$\sum x_i y_i = 12{,}240; \quad \sum x_i^2 = 462; \quad \sum y_i^2 = 335{,}975.$$

(i) State what is meant by response and explanatory variables in a linear regression



(ii) Comment on the relationship between school quality index and house price, using the plot.

Pearson's correlation coefficient between the data is given as r = 0.7.

(iii) A statistical test is performed, using Fisher's transformation, to determine whether Pearson's population correlation coefficient is significantly different from zero, i.e. for H0: $\rho = 0$ vs H1: $\rho \neq 0$.

(a) Identify which one of the following options gives the correct value of the test statistic for this test:

A1 2.295

A2 6.071

A3 2.743

A4 4.009

(b) Write down the conclusion of the test at the 5% level of significance, including the relevant critical value(s) from the Actuarial Formulae and Tables.

Unit 4

PRACTICE QUESTION

The linear regression line, of house prices (y) on school index (x), is given as
$\hat{y} = 133.8 + 7.386x$.

(iv) A t test is performed to determine if the slope parameter is significantly different from 0.

(a) Identify which one of the following options gives the correct values of the sums S_{xx}, S_{yy} , S_{xy} for the house prices (y) and school index (x) data:

A1 $S_{xx} = 32.8$; $S_{yy} = 2,415.4$; $S_{xy} = 235$

A2 S_{xx} = 20.5; S_{yy} = 3,131.2; S_{xy} = 182

A3 S_{xx} = 26.4; S_{yy} = 2,912.5; S_{xy} = 195

A4 S_{xx} = 35.2; S_{yy} = 2,817.4; S_{xy} = 247

(b) Calculate the value of the test statistic.

(c) Write down the distribution of the test statistic, if the null hypothesis of the test is correct.

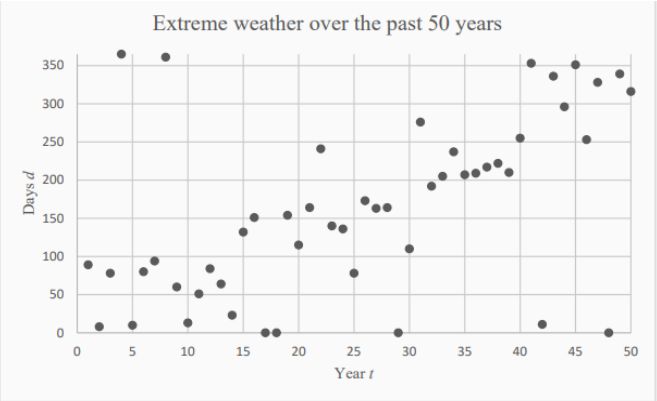
(d) Write down the conclusion of the test at the 5% level of significance, including the relevant critical value(s) from the Actuarial Formulae and Tables.

(v) Comment on the results in parts (iii)(b) and (iv)(d).

15. CS1A April 2021 Question 8

An initial investigation into climate change has been conducted using climate change data from the past 50 years, collected by the International Meteorological Society. For each year, t, the number of consecutive days, d, of extreme weather was recorded. The total number of days in any year is 365 and extreme weather is defined as a rainless day with temperatures in excess of 28 degrees Celsius.

An Actuary has performed a preliminary statistical analysis on the data. Below is a scatter plot of the Actuary's findings:



The Actuary also fitted a least squares regression line for extreme weather days on year, giving:
$\hat{d}$ = 147.39 – 5.82601t, and calculated the coefficient of determination for this regression line as:
R^2 = 91.5%

(i) Comment on the plot and the Actuary's analysis.

A separate analysis, on the same data, is undertaken independently by a statistician. Below are the key summaries of their analysis:

$$\sum t = 1{,}275 \quad \sum t^2 = 42{,}925 \quad \sum d = 8{,}502 \quad \sum d^2 = 1{,}911{,}378 \quad \sum td = 282{,}724$$

(ii) Verify that the equation of the statistician's least squares fitted regression line of extreme weather days on year is given by:
$\hat{d}$ = 8.59592 + 6.33114t.

(iii) (a) Determine the standard error of the estimated slope coefficient in part (ii).

(b) Test the null hypothesis of 'no linear relationship' at the 1% confidence level, using the equation in part (ii).

(c) Determine a 99% confidence interval for the underlying slope coefficient for the linear model, using the equation in part (ii).

Further climate change data are collected from an alternative independent data source, also covering the past 50 years. These data were analysed and resulted in an estimated slope coefficient of:
$\hat{\beta}$ = 5.21456 with standard error 1.98276

Unit 4

PRACTICE QUESTION

(iv) (a) Test the 'no linear relationship' hypothesis at the 1% confidence level based on the further climate change data.
(b) Determine a 99% confidence interval for the underlying slope coefficient β based on the alternative climate change data.

(v) Comment on whether or not the underlying slope coefficients, for the statistician's data in part (ii) and the independent data in part (iv), can be regarded as being equal.

(vi) Discuss why the results of the tests in parts (iii)(b) and (iv)(a) seem to contradict the conclusion in part (v).

16. CS1A September 2021 Question 9.

An actuarial analyst working in an investment bank believes that a firm's first year percentage return (y) depends on its revenues (x).

The table below provides a summary of x, y and the natural logarithmic revenue (z) for 110 firms.

	Mean	*Median*	*Sample standard deviation*	*Minimum*	*Maximum*
y	0.106	–0.130	0.824	–0.938	4.333
x (£ million)	134.487	39.971	261.881	0.099	1455.761
$z = \log(x)$	3.686	3.688	1.698	–2.316	7.283

The analyst determined that the correlation between y and x is −0.0175 and that the linear regression line of the return on the revenue is

$$\hat{y} = \hat{a} + \hat{b}x.$$

(i) (a) Identify which one of the following options gives the correct values of the coefficient estimates $\hat{a}$ and $\hat{b}$:

A $\hat{a} = 0.113$ and $\hat{b} = -5.506 \times 10^{-5}$

B $\hat{a} = -5.506 \times 10^{-5}$ and $\hat{b} = 0.113$

C $\hat{a} = 748.1227$ and $\hat{b} = -5.562$

D $\hat{a} = -5.562$ and $\hat{b} = 748.1227$

(b) Calculate the fitted return for a firm with revenue 95.55.

The analyst estimated the regression using the logarithm revenues (z) and y as

$$\hat{y} = 0.438 - 0.090z$$

(ii) (a) Calculate the fitted return for the firm with revenue 95.55 (£ million) using the regression model with the logarithmic revenues.
(b) Comment on the result in parts (ii)(a) and (i)(b).
(c) Calculate the value of the sum S_{zy}.

(iii) Perform a statistical test at the 10% significance level to determine if the logarithmic revenues significantly affect the percentage returns.

The analyst speculated that, other things being equal, firms with greater revenues will be more stable and thus enjoy a larger return. They considered the null hypothesis of no relation between z and y.

(iv) Perform a statistical test at the 10% significance level to determine whether the analyst's speculation is correct. Your answer should include the hypotheses of the test.

(v) Calculate Pearson's correlation coefficient between z and y.

A client is considering investing in a firm that has z = 2.
(vi) (a) Calculate the client's predicted first year percentage return.
(b) Calculate an approximate 95% confidence interval corresponding to the predicted percentage return in part (vi)(a).

A firm in the data has logarithmic revenue z = 1.76 and the highest first year percentage return y = 4.333.
(vii) (a) Calculate the residual for this observation.
(b) Comment on the observed data for this firm using part (vii)(a).

17. CS1A April 2022 Question 9.

Consider the linear regression model in which the response variable Y_i is linked to the explanatory variable X_i by the following equation:

$$Y_i = \alpha + \beta X_i + e_i, \; i = 1, \ldots, n,$$

where e_i are the error terms and data $(x_i, y_i), i = 1, \ldots, n$, are available.

(i) Comment on whether or not the linear regression model as presented above can be used to make inferences on parameters α and β.

Unit 4

PRACTICE QUESTION

The coefficient of determination for this model is given by $R^2 = \frac{s_{xy}^2}{S_{xx} S_y}$.

(ii) Verify that R^2 gives the proportion of the total variability of Y 'explained' by the linear regression model.

Consider the multiple linear regression model where the response variable Y_i is related to explanatory variables $X_1, X_2, ..., X_k$ by:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + e_i, \; i = 1, ..., n,$$

where e_i are the error terms and relevant data are available.

(iii) Suggest three ways for assessing the fit of the multiple linear regression model to a set of data.

A forward selection process is used for selecting explanatory variables in the multiple linear regression model.

(iv) Explain whether the coefficient of determination, R^2, can be used as a criterion for selecting variables when applying this process.

A multiple linear regression model with four explanatory variables (X_1, X_2, X_3, X_4) is fitted to a set of data, and a forward selection process is used for selecting the optimal set of explanatory variables.

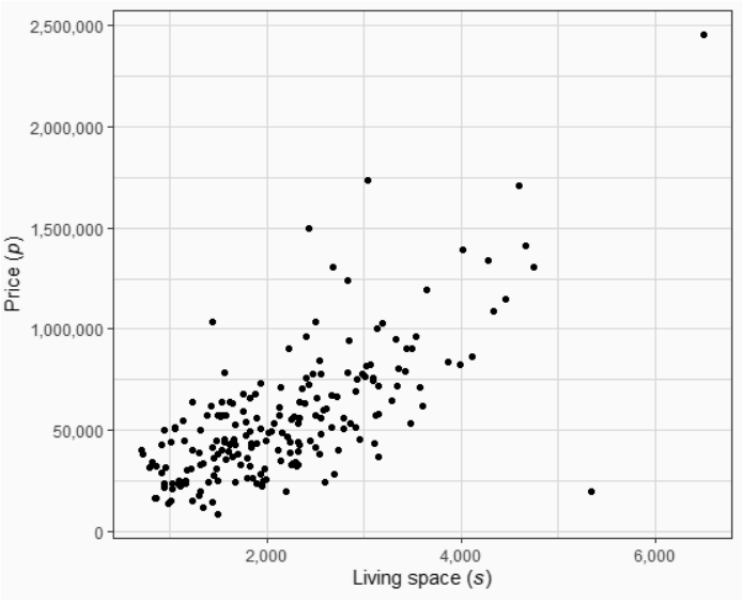
Some output of this process is shown in the following table:

Model	R^2	Adjusted R^2
X_1	0.7322	0.7167
$X_1 + X_4$	0.8018	0.7712
$X_1 + X_4 + X_3$	0.8253	0.7805
$X_1 + X_4 + X_3 + X_2$	0.8259	0.7684

(v) Determine the optimal set of explanatory variables using this output.

18. CS1A September 2022 Question 9

A Banking Analyst believes that living space, s, measured in square feet, is a good predictor of the price, p, of a property. The Analyst produces the figure below using a sample of 200 properties collected in a big city.



(i) Comment on the graph.

The Banking Analyst fits a least squares regression line for the logarithmic price (y = ln(p)) of the properties on the logarithm of the living space (x = ln(s)), using the summary of x and y shown below:

$$\sum x = 1,519.632\,;\ \sum x^2 = 11,583.92\,;\ \sum y = 2,616.206\,;\ \sum y^2 = 34,283.44$$

$$\sum yx = 19,908.94;\ \bar{y} = 13.081;\ \bar{x} = 7.598$$

(ii) Determine the Banking Analyst's least squares fitted regression line.

(iii) Calculate the coefficient of determination for the regression line determined in part (ii).

(iv) Calculate a two-sided 95% confidence interval for β, the slope of the true regression line.

(v) Test the hypotheses H0: β = 1 vs H1: β ≠ 1 at the 5% significance level.

(vi) Determine the 95% confidence interval for the expected price of a property with 1,930 square feet of living space.

Unit 4

PRACTICE QUESTION

(vii) Determine the 95% prediction interval for the price of a property with 1,930 square feet of living space.

(viii) Comment on your answer to parts (vi) and (vii).

The Banking Analyst fitted another least squares regression line for the price of the properties, depending on the square feet of living space and also the year the property was built. The coefficient of determination for this regression line is R2 = 60%.

(ix) Comment on the result from this second regression line and your answer to part (iii).

Generalized linear models

1. CT6 April 2012 Question 1

(i) Define what it means for a random variable to belong to an exponential family.

(ii) Show that if a random variable has the exponential distribution it belongs to an exponential family.

2. CT6 September 2013 Question 8

The number of claims per month Y arising on a certain portfolio of insurance policies is to be modelled using a modified geometric distribution with probability density given by:

$$p(\alpha) = \left(\frac{\alpha^{y-1}}{(1+\alpha)^{y}}\right);\ y = 1, 2, 3$$

where a is an unknown positive parameter. The most recent four months have resulted in claim numbers of 8, 6, 10 and 9.

(i) Derive the maximum likelihood estimate of α.

(ii) Show that Y belongs to an exponential family of distributions and suggest its natural parameter.

5. CT6 April 2014 Question 10

For a certain portfolio of insurance policies the number of claims on the i^{th} policy in the j^{th} year of cover is denoted by Y_{ij}. The distribution of Y_{ij} is given by:

Unit 4

PRACTICE QUESTION

$$P(Y_{ij}) = \theta_{ij}(1-\theta_{ij})^{y} \qquad y = 0,1,2,\ldots$$

where $0 \le \theta_{ij} \le 1$ are unknown parameters with $i = 1,2,\ldots,k$ and $j = 1,2,\ldots,l$.

(i) Derive the maximum likelihood estimate of θ_{ij} given the single observed data point y_{ij}

(ii) Write $P(Y_{ij} = y_{ij})$ in exponential family form and specify the parameters.

(iii) Describe the different characteristics of Pearson and deviance residuals.

6. CT6 September 2014 Question 2

(i) List the three main components of a generalised linear model.

(ii) Explain what is meant by a saturated model and discuss whether such a model is useful in practice.

7. CT6 October 2015 Question 6

(i) Explain what is meant by a saturated model.

(ii) State the definition of the scaled deviance in a fitting under generalized linear modelling.

(iii) (a) Define both Pearson and deviance residuals.
(b) Explain how these two types of residuals are generally different.
(c) State in which case they are the same.

9. CT6 April 2016 Question 10

(i) State the general expression of the exponential families of distributions and use this to derive the relevant expressions for the mean and the variance of these distributions.

(ii) Extend the result in (i) to obtain an expression for the third central moment.

(iii) Show that the following density function belongs to the exponential family of distributions:

Unit 4

PRACTICE QUESTION

$$f(x)=\frac{\alpha^{\alpha}}{\mu^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x\frac{\alpha}{\mu}}$$

(iv) Using the results in (i) and (ii) obtain the second and third central moments for this distribution.

10. CT6 September 2016 Question 6

Assume that the numbers of accidents for three different risks in five years are as follows:

	Year 1	*Year 2*	*Year 3*	*Year 4*	*Year 5*	***Total***
Risk A	1	4	5	0	2	**12**
Risk B	1	6	4	6	5	**22**
Risk C	5	6	4	9	4	**28**

An actuary is modelling each risk according to a Poisson distribution.

(i) Determine the Poisson parameter for each risk using the method of maximum likelihood estimation.

(ii) Test the hypothesis that the three risks have the same claim rate, using the scaled deviances.

11. CT6 April 2017 Question 5

(i) Show that the following discrete distribution belongs to the exponential family of distributions.

$$f(y;\mu)=\binom{n}{ny}\mu^{ny}(1-\mu)^{n-ny} \qquad y=0,\frac{1}{n},\frac{2}{n},\ldots,1$$

(ii) Derive expressions for the mean and variance of the distribution, E(Y) and var(Y), using your answer to part (i).

12. CT6 September 2017 Question 7

A random variable X follows a Poisson distribution with parameter l.

(i) Show that the distribution of X is a member of the exponential family of distributions.

(ii) Show that the mean of X equals the variance of X, using your answer to part (i).

(iii) Describe the three key components required when fitting a Generalised Linear Model (GLM).

Unit 4

PRACTICE QUESTION

13. CS1A September 2020 Question 5

An insurance portfolio has a set of n policies (i = 1, 2, …, n), for which the company has recorded the number of claims per month, Y_{ij}, for m months (j = 1, 2, …, m). It is assumed that the number of claims for each policy, for each month, are independent Poisson random variables with $E[Y_{ij}] = \mu_{ij}$. These random variables are modelled using a simple generalized linear model, with $\log(\mu_{ij}) = \beta_i$ for (i = 1, 2, …, n).

(i) Derive the maximum likelihood estimator of β_i

(ii) Show that the deviance for this model is:

$$D = 2\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{y_{ij}\log log\left(\frac{y_{ij}}{\bar{y}_i}\right) - \left(y_{ij} - \bar{y}_i\right)\right\}$$

where $\bar{y}_i$ is the average number of claims per month for policy i:

$$\bar{y}_i = \sum_{j=1}^{m}\frac{y_{ij}}{m}$$

The company has data for each month over a three-year period. For one policy, the average number of claims per month was 18.95. In the most recent month for this policy, there were seven claims.

(iii) Determine the part of the total deviance that comes from this single observation.

14. CS1A September 2020 Question 7

The probability density function of a Normal distribution is given as follows:

$$f\left(x, m, s^2\right) = \frac{1}{s\sqrt{2\pi}}exp\left(-\frac{1}{2s^2}(x - m)^2\right)$$

with $-\infty < x < \infty, -\infty < m < \infty, s > 0$.

(i) Identify which one of the following options gives the correct expression for the exponential family of the density f.

A1 $\frac{1}{\sqrt{2\pi}}exp\left(\frac{xm-\frac{m^2}{2}}{s^2} - \frac{x^2}{2s^2} - \ln ln\, s\right)$

A2 $exp\left(\frac{xm-\frac{m^2}{2}}{s^2} - \frac{x^2}{2s^2} - \frac{\ln ln\left(2\pi s^2\right)}{2}\right)$

A3 $exp\left(\frac{x(2m-x)}{2s^2} - \frac{\frac{m^2}{2}}{s^2} - \frac{\ln ln\left(2\pi s^2\right)}{2}\right)$

A4 $exp\left(\frac{1}{s^2}\left(xm - \frac{m^2}{2} - \frac{x^2}{2}\right) - \frac{\ln ln\left(2\pi s^2\right)}{2}\right)$

Unit 4

PRACTICE QUESTION

(ii) Identify which one of the following options gives the natural parameter θ, the scale parameter ϕ, and the relevant functions $b(\theta), a(\phi)$ and $c(x, \phi)$ of the exponential family for this distribution, using your answer to part (i).

A1 $\theta = m, \phi = s^2, b(\theta) = m^2, a(\phi) = \frac{s^2}{2}, c(x, \phi) = -\frac{1}{2}\left(x^2 + ln\left(2\pi s^2\right)\right)$

A2 $\theta = m, \phi = \frac{s^2}{2}, b(\theta) = m^2, a(\phi) = \frac{s^2}{2}, c(x, \phi) = -\frac{1}{2}\left(\frac{x^2}{s^s} + ln\left(2\pi s^2\right)\right)$

A3 $\theta = s^2, \phi = m, b(\theta) = m^2, a(\phi) = \frac{s^2}{2}, c(x, \phi) = -\frac{1}{2}\left(x^2 + \frac{ln\left(2\pi x^2\right)}{2}\right)$

A4 $\theta = m, \phi = s^2, b(\theta) = \frac{m^2}{2}, a(\phi) = s^2, c(x, \phi) = -\frac{1}{2}\left(\frac{x^2}{s^2} + ln\left(2\pi s^2\right)\right)$

An analyst found that the mean and standard deviation of this distribution are $E(X) = m$ and $SD(X) = s^2$. In your answer you may denote θ by theta and ϕ by phi.

(iii) Justify, using the properties of the exponential family, whether or not the analyst is right about the mean and standard deviation of this distribution.

(iv) Contrast a numerical variable and a factor covariate in the context of a generalised linear model.

15. CS1A September 2020 Question 6

(i) State the three components of a Generalised Linear Model (GLM).

In a mortality model, the number of deaths Dx at age x is modelled with a GLM. Dx is assumed to have a Poisson distribution with expectation mx = exp(a + bx) for each age x, such that Dx ~ Poisson(exp(a + bx)).

(ii) State the specific form of each of the three components of the GLM for the above mortality model.

(iii) Identify which one of the following expressions gives the correct likelihood function as a function of the unknown parameters a and b based on the observed number of deaths for all ages 20 to 80 given by d20,..., d80, assuming that the numbers of deaths at different ages are independent.

A1 $$L(a,b) = \prod_{x=20}^{80} P[D_x = d_x] = \prod_{x=20}^{80} \frac{1}{d_x!} e^{-e^{(a+bx)}} e^{(a+bx)d_x}$$

A2 $$L(a,b) = \prod_{x=20}^{80} P[D_x = d_x] = \prod_{x=20}^{80} e^{e^{(a+bx)}} e^{(a+bx)d_x}$$

A3 $$L(a,b) = \prod_{x=20}^{80} P[D_x = d_x] = \prod_{x=20}^{80} \frac{1}{d_x!} e^{-e^{(a-bx)}} e^{(a-bx)d_x}$$

A4 $$L(a,b) = \prod_{x=20}^{80} P[D_x = d_x] = \prod_{x=20}^{80} \frac{1}{d_x!} e^{e^{(a+bx)d_x}} e^{-(a+bx)}$$

An analyst is reviewing the mortality model and is considering deaths only for ages between 40 to 43 inclusive.

The analyst collects data for deaths and estimates the parameters for a and b as follows:

$$d_{40} = 2; d_{41} = 3; d_{42} = 1; d_{43} = 0; a = 0.01512; b = -0.00686$$

(iv) Identify, using your answer to part (iii), which one of the following options gives the correct value of the likelihood function, based on the analyst's data and parameter estimates.

A1 0.00222

A2 4.05473

A3 0.0008

A4 4.32729

16. CS1A September 2021 Question 8

The number of hospital admissions for respiratory conditions in a big city was recorded over 150 days. The level of the concentration of a certain pollutant was also recorded ('low', 'medium', 'high'), together with the mean temperature (in degrees Celsius) on the day. Part of the data is shown below.

A generalized linear model is to be fitted to investigate the dependence of the number of hospital admissions on mean temperature and pollutant concentration.

Unit 4

PRACTICE QUESTION

(i) Write down a suitable model for the number of hospital admissions.

(ii) Justify the inclusion of the terms that you have used in the linear predictor in part (i).

A statistician fitted a GLM, and obtained the following summary:

Coefficients:	Estimate	Std. error	z value	$Pr(>\|z\|)$
(Intercept)	-0.372	0.053	-6.916	$4.66e-12$ ***
X_1	0.090	0.015	5.676	$1.38e-08$ ***
X_2 Medium	-0.100	0.080	-1.244	0.213570
X_2 High	0.298	0.082	3.614	0.000301 ***
$X_1 : X_2$ Medium	0.036	0.023	1.551	0.120933
$X_1 : X_2$ High	-0.076	0.028	-2.705	0.006825 **

Suppose that, on a different day, the pollutant concentration is High and the mean temperature is 19 degrees Celsius.

(iii) Write down the linear function of the parameters the statistician should use in constructing a predictor of the number of hospital admissions on that day.

(iv) Explain why estimates for X_2 Low and $X_1 : X_2$ Low are not shown in the summary of the results above.

(v) Comment on the impact of the pollutant concentration on the number of hospital admissions, based on the summary of results above.

17. CS1A April 2022 Question 7

The probability density function of a gamma distribution is parameterised as follows:

Unit 4

PRACTICE QUESTION

$$f(x) = \frac{\left(\frac{\mu}{\sigma^2}\right)^{(\mu^2/\sigma^2)}}{\Gamma\left(\frac{\mu^2}{\sigma^2}\right)} x^{\left(\frac{\mu^2}{\sigma^2}\right)-1} e^{-x\mu/\sigma^2}, \quad x \geq 0, \quad \mu, \sigma > 0.$$

This density can be expressed in the form of the exponential family, as follows:

$$\theta = -\frac{1}{\mu}, \quad b(\theta) = -\log(-\theta), \quad \phi = \frac{\mu^2}{\sigma^2}, \quad \alpha(\phi) = \frac{1}{\phi},$$

$$c(x, \phi) = (\phi - 1)\log x - \log\Gamma(\phi) + \phi \log \phi,$$

where the exponential family notation is the same as that in the Actuarial Formulae and Tables book.

(i) Justify that μ and σ^2 are the mean and the variance of the distribution, respectively, using the properties of the exponential family.

An actuary is modelling the relationship between claim size and the time spent processing the claim, called operational time (opt). A statistician suggests using a model with the claim size being the response variable following the gamma distribution given above.

(ii) Comment on why a gamma distribution may be more suitable than the Normal distribution for the claim sizes.

The actuary decided to fit a generalised linear model (GLM) with a gamma family and obtained the following estimates:

Parameters:

	Estimate	Standard error
Intercept	7.51621	0.03310
opt	0.06084	0.00296

(iii) Explain, using the model output shown above, whether the variable 'opt' is significant or not.

Another statistician has suggested that an alternative model needs to take into account a legal representation variable, which shows whether or not an insured person has legal representation.

(iv) Explain the difference between the variables 'opt' and 'legal representation' in a statistical sense in the context of a GLM.

The actuary now has to choose between the following two models for the claim size:

Model 1: Only opt is used as a covariate.
Model 2: Both opt and legal representation are used as covariates.

An analysis of variance (ANOVA) was carried out to assess the significance of the two covariates: opt and legal representation (denoted by lr). The results obtained are given below, where claim size is denoted by cs:
Model 1: cs = 7.52 + 0.06 × opt
Model 2: cs = 3.6 + 0.04 × opt + 2.32 × l_r

	Resid. df	*Resid. dev*	*Df*	*Deviance*	*Pr(>Chi)*
Model 1	45	39.987			
Model 2	44	15.869	1	24.118	0.000286

(v) Determine which model provides the better fit to the data.

18. CS1A September 2022 Question 5

The claim amounts in an insurance company's car insurance portfolio follow a gamma distribution. The company is modelling the claims it receives and is considering a Generalized Linear Model (GLM), with claim amounts as the response variable and four relevant covariates:

- The age (x) of the policyholder
- The experience of the policyholder (a category between 1 and 4, based on the number of years of driving experience)
- The gender of the policyholder (1 = male, 2 = female)
- The car insurance group (a rating between 1 and 20, indicating the level of risk).

(i) State the form of the linear predictor of the GLM when all the covariates are included in the model as main effects.

(ii) Explain all the terms used in the linear predictor in your answer to part (i).

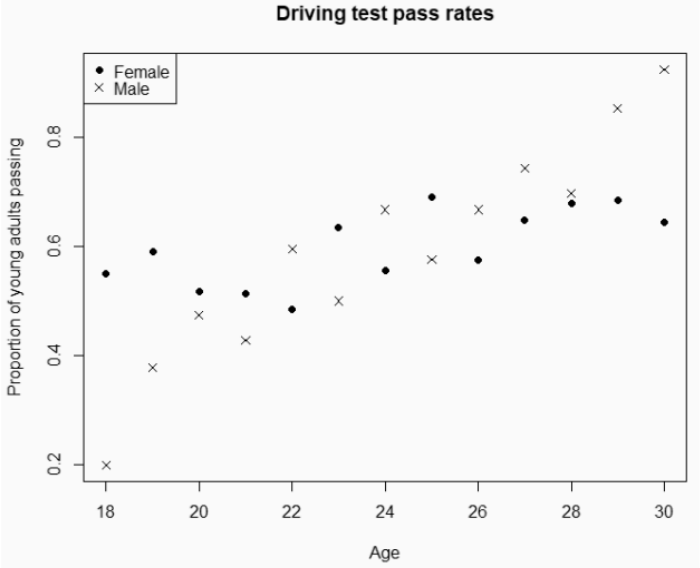
(iii) State how the linear predictor in your answer to part (i) changes if an interaction between the covariates showing policyholder age and car insurance group is also included in the model.

You should explain all the terms used in the new linear predictor.
The company is considering whether to include the interaction term between policyholder age and car insurance group. The scaled deviance of the GLM without the interaction term in part (i) has been calculated as 422.5. For the GLM including the interaction in part (iii), the scaled deviance is equal to 310.3.

(iv) Compare the two models by performing a suitable test for investigating whether the model including the interaction term is a significant improvement over the model without the interaction term.

20. CS1A September 2022 Question 3

A study is undertaken in order to devise a model to predict the probabilities of young adults passing a driving test. The data was collected on the basis of results over a 30-day period. An Analyst's observations for any given gender and age group are of the form Y/n, where Y is the number passing the test and n is the number taking the test. The Analyst plots the proportion of young adults passing by age for males and females as shown below.



(i) Comment on the graph.

The Analyst believes that age and gender are variables that influence whether or not a person will pass a driving test. The Analyst fitted a Generalised Linear Model (GLM), with a canonical link function, to investigate such an influence by including the interaction term between the two explanatory variables.

(ii) Write down a suitable model for the proportion passing the test.

The summary of the fitted model is provided in the form of linear predictors for females (F) and males (M) respectively as:

$$\hat{\eta}_F = -0.968 + (0.056) \times Age \text{ and } \hat{\eta}_M = -4.584 + (0.209) \times Age$$

(iii) Determine the proportion of 22-year-old females predicted by the model to pass the test.

Unit 4

PRACTICE QUESTION

Using the fitted GLM model, the Analyst derives the following expression for the ratio of the probability of passing the test (μ) over the probability of failing (1 – μ) for males:

$$\frac{\hat{\mu}}{1-\hat{\mu}} = \exp exp\left(\hat{\eta}_M\right) = \exp exp\left(-4.584 + 0.209 \times Age\right)$$

(iv) Comment on this expression with respect to the probability of passing the test.