IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Probability and Statistics

Chapter: Unit 4

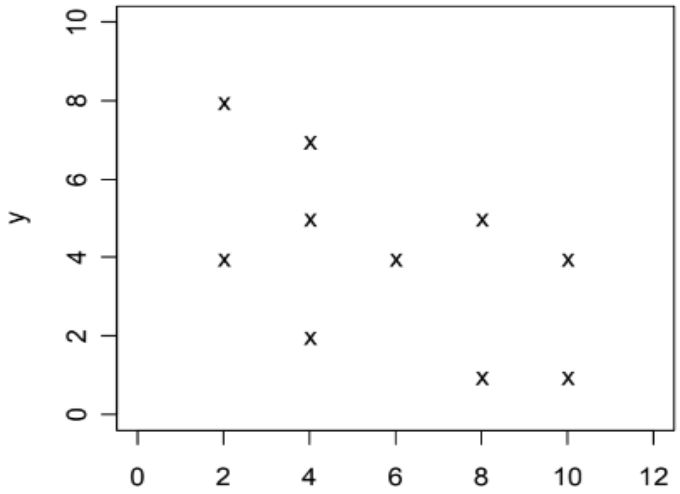
Category: Practice question

Part 1

- **Correlation analysis and regression**

1. CT3 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.

Unit 4

PRACTICE QUESTION

Ans:

(i) $R^2 = 0.3135$ (or 31.4%)
(ii) $y = 6.5837 - 0.4282x$
(iii) (–0.945, 0.088)
(iv) 2.730
(v) 0.966
(vi) Addition of new observation makes data more randomly scattered. The strength of the linear relationship is reduced from "weak" to "almost nothing".

2. CT3 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:

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).

Ans:

(i) r = 0.796084
(ii) p-value > 0.49, No evidence against the null hypothesis
(iii) $\hat{a}$= 130.625, $\hat{b}$= 8.125
(iv) R^2 = 0.634
(v) Any increase in school quality by 1 index-point, leads to an increase of 8.125 in the house price index.

3. CT3 October 2012 Question 13

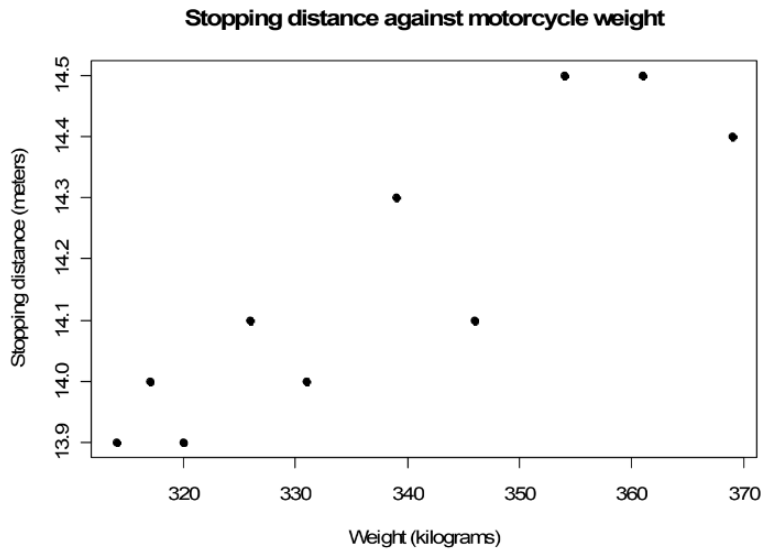
The following data give the weight, in kilograms, of a random sample of 10 different models of similar motorcycles and the distance, in meters, required to stop from a speed of 20 miles per hour.

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).

Ans:

(i) (a) The scatter plot suggests a positive linear association between weight and stopping distance.

(b) r = 0.892

(ii) P-value = $\Pr(Z \geq 3.79) \approx 0.0001$, so there is very strong evidence against H_0 and we conclude that motorcycle weight and stopping distance are positively correlated.

(iii) (a) $\hat{y} = 10.48 + 0.01092x$

(b) $R^2 = 0.7956$

Unit 4

PRACTICE QUESTION

(c) For every additional unit (kilogram) of weight the stopping distance is expected to increase by $\hat{\beta}$= 0.01092 metres. So, for 10 kilograms of weight the distance is expected to increase by 0.109 meters.

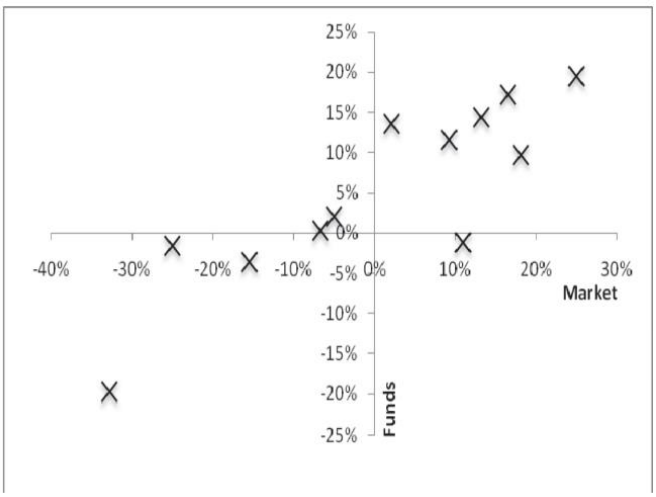
3. CT3 October 2013 Question 10

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.

Unit 4

PRACTICE QUESTION

(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).

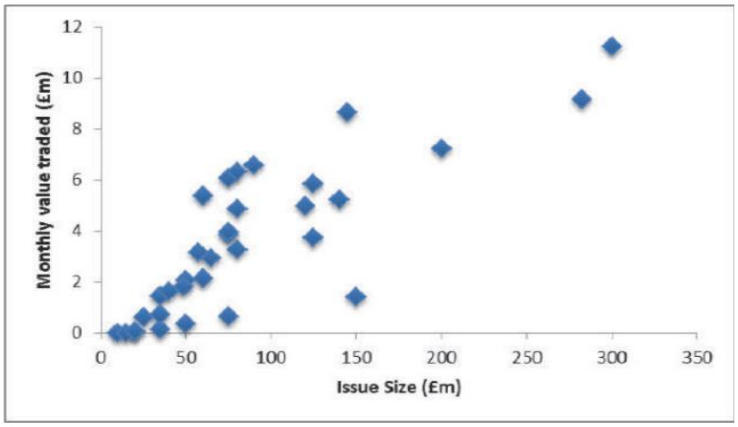
(v) Comment on your answers to parts (ii)(b) and (iv).

Ans:

(i) There is a positive linear relationship between the two.
(ii) (a) r = 0.867
(b) test stat = 5.50, Critical value = 3.169,
So reject H_0 that correlation coefficient = 0 at 1% level
(iii) $\hat{a}$ = 0.0473, $\hat{b}$ = 0.539
(iv) (0.321,0.757)
(v) C.I. does not contain zero. Consistent with correlation coefficient not equal to zero as the test is actually the same. Both suggest that the hedge industry's claim that correlation is low may not be correct.

4. CT3 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.



Unit 4

PRACTICE QUESTION

(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.

Ans:

(i) There appears to be a positive linear relationship

(ii) (a) r = 0.8294

(b) test stat = 8.266, At 0.5% level t_{31} = 2.744 which < test statistic. So, reject H_0.

(iii) $v_i = 0.398 + 0.0359 s_i$

(iv) Testing whether β is significantly different from zero is mathematically the same as testing whether the correlation coefficient is significantly different from zero. As H_0 was rejected in (ii)(b), we can conclude testing H_0: $\beta = 0$ would give the same result.

(v) It is true that extreme observations can determine the strength of a linear relationship. However, there are many more bonds in the central part of the data and we would consequently expect a greater range of value traded.

5. CT3 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, y_i	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).

Ans:

(i) r = 0.76153

(ii) (0.122688, 0.95417)

(iii) $\hat{a}$ = 1.29891, $\hat{b}$ = 0.293137

(iv) R^2 = 0.58

(v) About 58% of the total variability of the response "cigarettes per day" is statistically explained by alcohol consumption.

Unit 4

PRACTICE QUESTION

6. CT3 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:

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

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

(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.

Ans:

(i) r = 0.1470326

(ii) test stat = 0.71, Critical value = +-1.714, Since this is a two-sided test and 0.71 is within the interval [–1.714, 1.714] the null hypothesis cannot be rejected at 10% level of significance.

7. CT3 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.

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

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

(iv) Estimate the parameters γ and β.

Ans:

(i) test stat =1.38, Critical value = 1.7, Therefore, there is no evidence (at 5% level) to suggest that the variance for new buildings is larger.
(ii) Test stat = -1.06, Critical value = between 1.98 and 2.00, There is no evidence to suggest that the mean maintenance costs of new buildings are different from mean maintenance costs of old buildings.
(iii) $\hat{\rho}$ (A, X) = 0.7
(iv) $\hat{\gamma}$ = 0.04511, $\hat{\beta}$ =1.59

Unit 4

PRACTICE QUESTION

8. CT3 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.

Ans:

(i) (a) r = -0.843

(b) There appears to be strong negative linear correlation between mileage and price.

(ii) $\hat{y} = 7.30 - 0.0469x$

(iii) (a) (−0.071, −0.022)

(b) Since the value zero is not included in the interval, the suggestion in

the article seems valid.
(iv) Estimated difference in price will be £234.34.

9. CT3 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

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

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

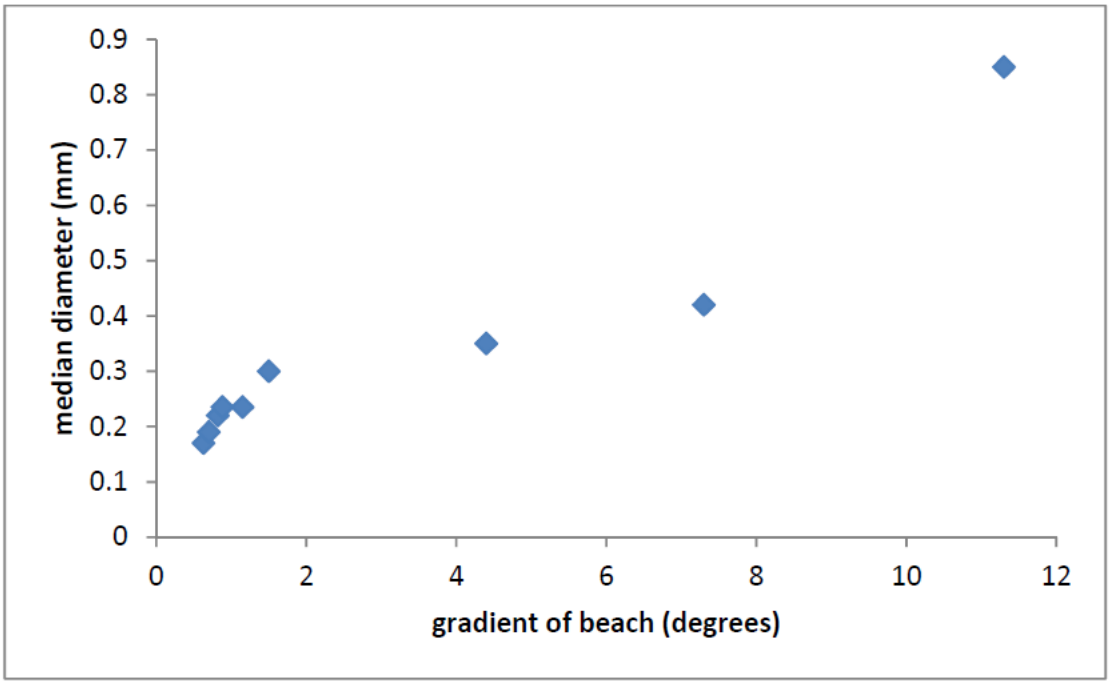
The geologist assumes that the error terms in the linear regression are normally distributed.
(ii) Perform a test to determine whether the slope coefficient is significantly different from zero.

(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.

Ans:

(i) $d = 0.1609 = 0.05306g$
(ii) test stat = 8.438, $t_{7;0.975} = 2.365$, so reject H_0: $\beta = 0$
(iii) (0.267,0.373)



(iv) (a)

(b) With only three observations for $g>1.5$, the slope is determined by a small amount of data. Getting more observations in that range would give a better analysis.

10. CT3 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 $i = 1,\ldots,9$.

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

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

Unit 4

PRACTICE QUESTION

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

(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.

Ans:

(i) In bivariate data, the response variable is a random variable whose value may be influenced by the value of the explanatory variable.
(ii) r = 0.6887
(iii) test stat = 2.071, Critical value =1.96, Therefore reject at 5% level H_0.
(iv) t =16.45 + 1.405y

11. CT3 September 2018 Question 10

A statistician has a series of bivariate data $\{(x_1,y_1), (x_2,y_2), \ldots (x_n,y_n)\}$ and wishes to perform a linear regression on these data.

(i) State the equation that must be minimised to give the least squares estimates of the regression coefficients.

(ii) Derive the least squares estimate of the slope coefficient from the equation in part (i).

For a sample of 44 fish, the age (days) and length (millimetres) of each fish are measured. Denote age by X and length by Y. The following summary data are obtained:

$\sum_{i=1}^{44} X_i$=3660, $\sum_{i=1}^{44} X_i^2$ = 389,684 $\sum_{i=1}^{44} Y_i$ = 136,727 $\sum_{i=1}^{44} Y_i^2$= 500,813,951 $\sum_{i=1}^{44} X_i Y_i$ = 13,609,918

(iii) Determine the coefficients for a linear regression of Y on X.
(iv) Calculate the sample correlation coefficient between x and y.

Unit 4

PRACTICE QUESTION

Ans:

(i) $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)\ ]^2$

(ii) $\hat{\beta} = \left(n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)\right) / (n\sum_{i=1}^{n} x_i^2 - [\sum_{i=1}^{n} x_i]^2)$

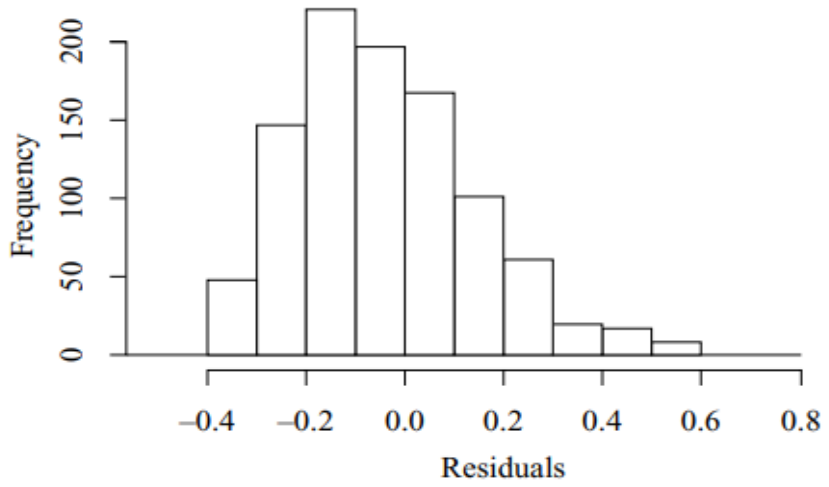
(iii) $\hat{\alpha}$ =924.68 $\hat{\beta}$ = 26.241

(iv) r = 0.879

12. 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

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

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

A linear model of the form y = α + βx + e is fitted to the data, where the error terms (e) independently follow a N(0, σ^2) distribution, and where a, b and s^2 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.

Ans:

(i) The regression slope suggests a positive relationship between the two variables, while the correlation coefficient shows a strong negative relationship.
(ii) The histogram suggests a non-symmetric distribution for the residuals. Non-symmetric about zero.
(iii) $\hat{y}=-1.066-4.013x$
(iv) (−45.879, −44.535)
(v) The width of the interval is only affected by $V(\hat{y})$, which depends on the new x value through the term $(x_{new}-\underline{x})^2$. This term will now be smaller as the new x_{new}=3.5 value is closer to $\underline{x}$ than x=11. Therefore, the interval will be narrower.

Unit 4

PRACTICE QUESTION

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.

(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.

A. A1 $y_i = a + bx_i$
B. $y_i = a + bx_i + z_i \; with \; E[z_i] = 0$
C. $y_i = a + bx_i + z_i \; with \; z_i \sim \chi^2, 299 \; df$
D. $y_i = a + bx_i + z_i \; with \; z_i \sim N(0, \sigma 2)$

Unit 4

PRACTICE QUESTION

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

Ans:

(i) Z_X has a chi-squared distribution with $n-1=299$ degrees of freedom

(ii) $E(Z_X)=299$, $V(Z_X)= 598$

(iii) A chi-squared distribution with 299 degrees of freedom is the distribution of a sum of 299 independent random variable that are all squared standard normally distributed. It follows from the CLT that a chi-squared distribution with a large number of degrees of freedom can be approximated with a normal distribution.

(iv) $q_{97.5}= 346.93$, $q_{2.5}= 251.07$

(v) [1814.12,1861.88]

(vi) [576.88,639.12]

(vii) [38370.22,53020.19]

(viii) Option D

(ix) $\hat{a} = -1152.268$, $\hat{b} = 0.9577082$

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}.

Unit 4

PRACTICE QUESTION

(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.

Ans:

(i) $S_{xx} = 25.9$, $S_{yy} = 16.1$, $S_{xy} = 19.3$

(ii) $r = 0.9451364$

(iii) $\hat{a} = 2.193$, $\hat{b} = 0.745$

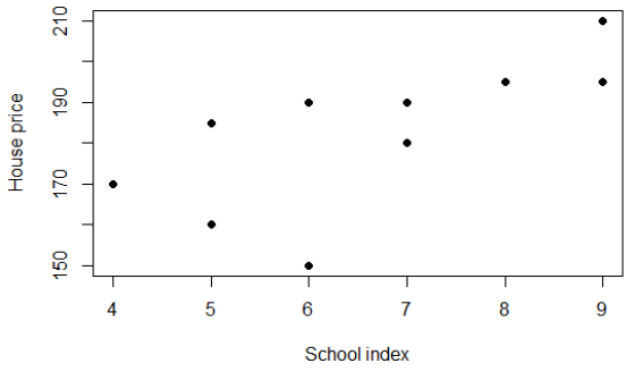
14. CS1A September 2020 Question 10

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.

Unit 4

PRACTICE QUESTION

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.

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).

Ans:

(i) In bivariate data, the response variable is a random variable whose value is influenced by the explanatory variable.

(ii) There is an increasing and relatively linear relationship. However the trend and linearity are not very clear around values x = 5, 6.

(iii) (a) Option A1

(b) This is a two-sided test with the 2.5% critical values being -1.96 and 1.96 So we reject H_0 at 5% significance level and conclude that Pearson's correlation coefficient is significantly different from zero.

(iv) (a) Option A3

(b) Test statistic = 2.80

(c) The test statistic follows a t-distribution with 8 df under the null hypothesis.

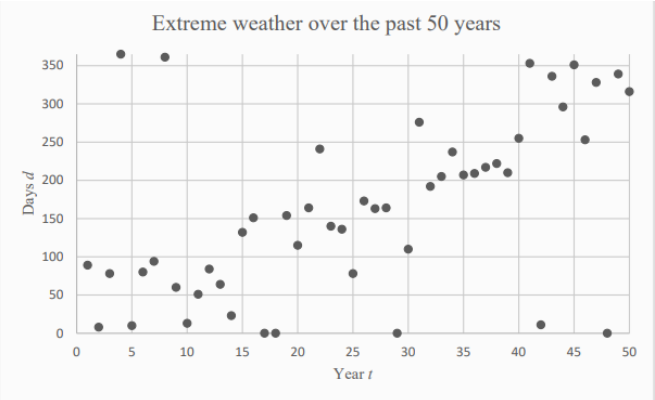
(d) This is a two-sided test with the 2.5% critical values being -2.306 and 2.306. We have evidence at 5% significance level to reject the null hypothesis that β=0.

(v) The two tests are actually similar therefore it is not surprising that they yield to the same conclusion that there is a linear relationship between house prices and school indices.

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).

Unit 4

PRACTICE QUESTION

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

(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).

Ans:

(i) There appears to be a number of possible outliers, (i.e. c0 or c365 days, these should be rechecked as they may be an error in the data or analysis.) The plot exhibits a strong positive linear relationship between days and year. R^2 percentage looks too high when compared to the scatterplot and the several outliers α value looks too high, we would expect it lower than 100 days, looking at the scatterplot β value sign looks to be the wrong way around, i.e. should be a positive. The number of days is bounded in the interval [0,366]. If the intention is to project into future years, it may have been better to fit a model that respects this restriction, e.g. do a logistic transformation on the number of days first (although the relationship may no longer be linear) (Can cover any 3 point for full marks)

(ii) $\hat{\alpha}$ = 8.596, $\hat{\beta}$ = 6.331

(iii) (a) 0.311

(b) test stat = 20.360, critical value= ±2.6832, there is strong evidence to reject H_0 at the 1% level, i.e. there is sufficient evidence to suggest that there is a strong linear relationship.

(c) (5.498 ,7.164)

(iv) (a) test stat = 2.630, CV = ±2.678, we have no evidence to reject H_0 at the 1% level. We conclude that there is insufficient evidence of a linear relationship

(b) (−0.095 ,10.524)

(v) The two confidence intervals overlap, with one being a subset of the other. This suggests that we confidently conclude that the underlying slope coefficients are

different. However, the large standard error leads to a wide confidence interval, meaning we lack evidence in the conclusion to the above bullet points.

(vi) The test conclusions in (iii)(b) and (iv)(a) appear to disagree. The test statistic in (iii)(b) lies well over the critical value whereas the test statistic in (iv)(a) lies just under the critical value. So this suggests that the slope coefficients may be different for the two sets of climate change data. Recording of past data, method of collection, errors in collection / the data etc from the alternative sources, treatment of outliers, differences in definition (e.g. location used) of extreme weather, may lead to the apparent differences observed

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).

Ans:
(i) (a) option A
(b) y* = 0.108
(ii) (a) y* = 0.028
(b) The return estimated with the log revenue is different from the return in part (i)(b) as expected
(c) S_{zz} = 314.269, S_{zy}= −28.284

Unit 4

PRACTICE QUESTION

(iii) test stat = −1.956, We have evidence at 10% significance level to reject the null hypothesis that β=0 and we conclude that the logarithmic revenues affect returns.
(iv) We do not have evidence to reject H_0 at 10% significance level. Firms with greater revenues do not necessary enjoy a larger return.
(v) r= −0.185
(vi) (a) $y*$= 0.258
(b) (−1.367 ,1.883)
(vii) (a) $\tilde{e}$= 4.053
(b) The residual is way above 0 and from the table the percentage return is 3 times the median.

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:

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

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

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

(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 $X_1, X_2, \dots, X_k$ by:

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

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.

Unit 4

PRACTICE QUESTION

(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

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

Ans:

(i) As it stands the model cannot be used for inference. We need further assumptions: the errors e_i are independent and $e_i \sim N(0,\sigma^2)$.

(ii) –

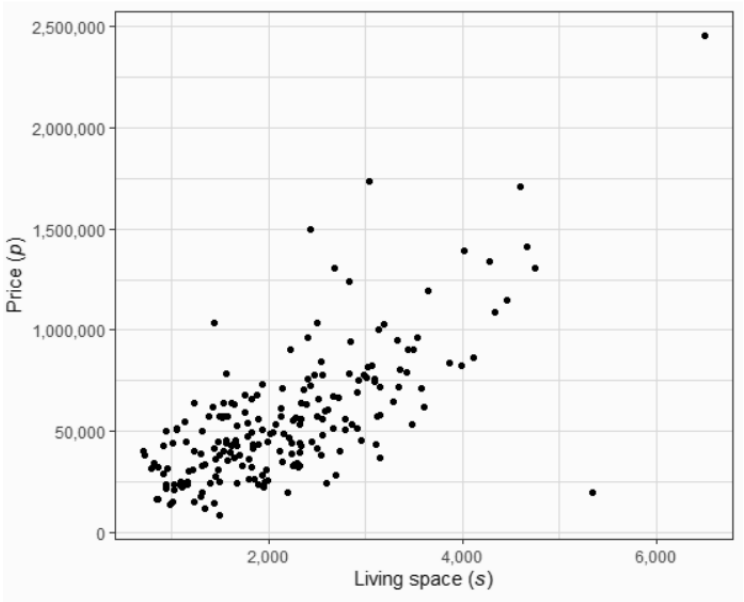
(iii) Possible approaches: Use R^2, Use adjusted $R2$, Plot residuals against fitted values (or explanatory variables)

(iv) R^2 cannot be used, Although increased values show a better fit of the model, it cannot decrease as more explanatory variables are added to the model.

(v) The adjusted R^2 should be used as a criterion. So, the model maximising the adjusted R^2 has explanatory variables X1+X4 + X3.

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:

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

$$\sum xy = 19{,}908.94;\ \underline{y} = 13.081;\ \underline{x} = 7.598$$

(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.

Unit 4

PRACTICE QUESTION

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

(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).

Ans:

(i) The plot exhibits a positive linear relationship between the price and the living space of a property however, There appear to be possible outliers (that could affect inferences).
(ii) $\hat{y}$=6.889+ 0.815x
(iii) R^2 = 0.41
(iv) (0.680, 0.950)
(v) The 95% two-sided confidence interval in (iv) does not contain the value 1, so the two-sided test conducted at the 5% level results in H0 being rejected.
(vi) (440647.3, 495835.7)
(vii) (202399.8, 1079490.8)
(viii) The variance for the predicted value is higher. Therefore, the CI is much wider (and contains the expected CI).
The predicted CI accounts for both the uncertainty in estimating the population mean and the random variation of the individual values.
(ix) The coefficient of determination for the simple linear regression model in (iii) is lower than that of the multiple linear regression model.
The year a property is built might be a good predictor for a property price.
Additional metrics e.g. AIC (OR adjusted R-Squared) should be used to confirm this.

19. CS1A April 2023 Question 10.

A Banking Analyst is assessing the performance of a newly developed credit risk model against experts' knowledge. The credit scores produced on a sample of twelve customers by the experts (x) and the model (y) are the following:

x	65.8	63.7	67.6	64.4	68.2	62.9	70.5	66.4	68.0	67.1	69.5	71.8
y	68.2	66.2	68.1	66.0	69.1	66.1	68.7	65.9	69.3	67.2	67.9	70.4

Summary statistics of the data are given below:

$$\sum x_i = 805.9 \quad \sum y_i = 813.1 \quad \sum x_i^2 = 54{,}203.21$$

$$\sum y_i^2 = 55{,}118.71 \quad \sum x_i y_i = 54{,}643.17$$

(i) Fit a linear regression line of y on x.

(ii) Calculate Pearson's correlation coefficient between the experts' and the Model's scores.

(iii) Perform a statistical test, using Fisher's transformation, to determine whether the population Pearson's correlation coefficient is significantly different from 0.8. Your answer should include the p-value of the test.

(iv) Construct a 99% confidence interval for the slope parameter of the linear regression line fitted in part (i)

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

The Analyst is informed that the scores on their own are not the most important aspect of the model. Instead, the performance of the model is assessed by how well it is able to predict the rank order of the twelve customers provided by the experts. A higher score corresponds to a better customer. The rankings of the customers based on their above scores are provided in the table below:

Rank (xi)	4	2	7	3	9	1	11	5	8	6	10	12
Rank (yi)	8	4	7	2	10	3	9	1	11	5	6	12

Unit 4

PRACTICE QUESTION

(vi) Calculate Spearman's rank correlation for the data between the model's and experts scores.
(vii) Comment on the model's alignment with the experts' opinion, based on your results from parts (ii) and (vi)

Ans:

(i) $\hat{y}$=37.067+0.457x
(ii) r = 0.830
(iii) W=1.188, 0.790
We have no evidence even at 1% against the null hypothesis that ρ=0.8
(iv) (0.150,0.764)
(v) In part (iv), the confidence interval does not contain 0 and we can conclude there is a linear relationship between the model and experts scores The test in part (iii) concluded no evidence against 0.8 correlation between the model's scores and experts' scores. The two results are consistent
(vi) 0.748
(vii) The Spearman correlation seems slightly lower than the Pearson's correlation but the model's alignment with the experts' opinion is good

20. CS1A September 2023 Question 9

The following nine pairs of data are given on observations from two random variables X and Y:

x	5	7	0	6	8	1	4	9	2
y	11.00	16.18	0.68	11.99	17.72	2.91	9.64	18.92	6.31

Summary statistics of the data are shown below:

$$\sum x_i = 42, \sum y_i = 95.35, \sum x_i^2 = 276, \sum y_i^2 = 1340.194, \sum x_i y_i = 606.33$$

(i) Perform a suitable statistical test to investigate the hypothesis that Pearson's population correlation coefficient for X and Y is positive, at significance level 0.01.

(ii) Comment on the relationship between X and Y, based on your answer in part (i).

(iii) Fit a simple linear regression model of y on x.

Unit 4

PRACTICE QUESTION

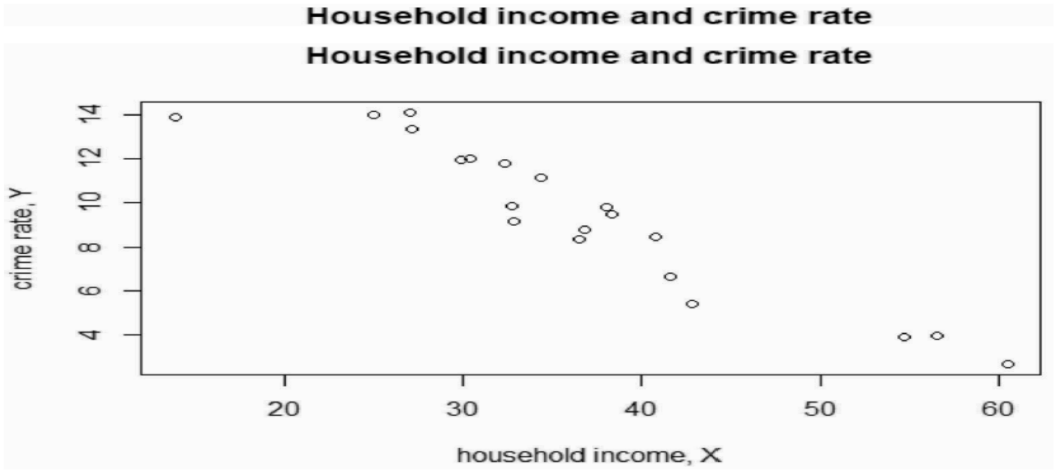
(iv) Determine a 99% confidence interval for the predicted mean response, when $x = 3$.

Ans:

(i) r = 0.993, test stat = 22.41, critical value = 2.998
We reject the null hypothesis in favour of the alternative that $\rho>0$.
(ii) The sample correlation and test suggest a strong positive linear relationship between X and Y.
(iii) $\hat{y}=1.182+2.017x$
(iv) (6.157, 8.309).

21. CS1A September 2023 Question 10

For a socio-economic analysis, a random sample of 20 regions is considered. An analyst collects data on average household income (in units of \$1,000, denoted by X) and crime rate (in percent, denoted by Y) for each region in the sample. The data are displayed in the following plot, and summary statistics are given below.



$$\sum_{i=1}^{20} x_i = 733, \sum_{i=1}^{20} y_i = 189, \sum_{i=1}^{20} x_i^2 = 29{,}203, \sum_{i=1}^{20} y_i^2 = 2{,}009, \sum_{i=1}^{20} x_i\, y_i = 6{,}208$$

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

(ii) Calculate Pearson's correlation coefficient for the relationship between household income and crime rate.

(iii) Justify why Pearson's correlation coefficient is appropriate in this context as compared to alternative correlation coefficient measures.

Unit 4

PRACTICE QUESTION

For the three regions with the highest household income the following data have been observed:

Household income (in units of \$1,000), X	54.73	56.61	60.54
Crime rate (%), Y	3.896	3.958	2.658

(iv) Perform a statistical test to decide if there is a significant difference in the mean crime rate between the group of regions with an average household income of more than \$50,000 and the group of regions with an average household income of less than \$50,000. You can assume that the variances of crime rates are the same in all regions.

(v) Comment on the assumption of equal variances made in part (iv).

Ans:

(i) [31.458,41.842]
(ii) r = −0.9955457
(iii) Pearson's correlation coefficient measures the strength of a linear relationship between two variables. Since we have two numerical variables using this coefficient is justified here. On the other hand, rank correlation coefficients are less appropriate since the numerical difference between two successive values has a meaning in this context, and numerical values therefore contain more information than just ranks.
(iv) Test stat = 4.79, critical value = 2.87844, The value of the test statistic is greater than the quantile, therefore reject null hypothesis of equal crime rates.
(v) The assumption seems to be unjustified as the spread of values for crime rate in the plot is very small for high income regions (50k) while it is rather wide for other regions. This raises questions about the validity of the test.

22. CS1A April 2024 Question 8

In a particular year, the members of a regional oil organisation decide to increase the production of oil. An analyst wishes to model the effect of this increase on the market price of oil and recorded the monthly average volume produced, x, in litres (l), and its price per litre, y (£) for eight consecutive months. The data are summarised as follows:

S_{xx}=3,535,237.5, $\sum_{i=1}^{8} y$= 181, $\sum_{i=1}^{8} y^2$= 181, S_{xy}=23,726.25

Unit 4

PRACTICE QUESTION

(i) Calculate Pearson's correlation coefficient between x and y.

(ii) Comment on your answer to part (i).

A second analyst thought that converting the volume to cubic metres (m^3), as $Z = x/1000$, would make the computation more efficient.

(iii) Write down S_{zz} expressed in terms of Sxx

(iv) Calculate the new Pearson's correlation coefficient between z and y.

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

(vi) Determine the least squares fitted regression line between y and z given that $\Sigma z = 0.16$.

(vii) Determine a 95% prediction interval for the price of oil, when the oil production is 1.5 m^3 based on the observed data.

Ans:

(i) r = −0.871
(ii) The correlation coefficient is negative, thus as the volume of the production increases the price decreases. The correlation is quite strong.
(iii) $S_{zz} = S_{xx} / 1000^2$
(iv) $r = -0.871$
(v) Parts (i) and (iv) give the same value. The Pearson correlation coefficient is invariant to this transformation.
(vi) $\hat{y} = -6.71z + 22.759$
(vii) (3.302, 22.080)

Part 2

- **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.

Ans:

(i) A random variable Y belongs to an exponential family if the pdf of Y can be written in the form

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right]$$

Where a, b and c are functions

(ii) –

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.

Ans:

(i) $\hat{\alpha}$=7.25

(ii) –

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:

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

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.

Ans:

(i) $\widehat{\theta_{ij}} = \frac{1}{1+y_{ij}}$

(ii) $P(Y_{ij} = y) = \exp\left[\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right]$

$\theta = log\,(1-\theta_{ij})$, $a(\varphi)=1$, $b(\theta)=-\log(1-e^{\theta})$, $c(y,\varphi)=0$

(iii) The Pearson residuals are often skewed for non normal data which makes the interpretation of residual plots difficult. Deviance residuals are usually more likely to be symmetrically distributed and are preferred for actuarial applications.

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.

Ans:

(i) The three main components are:
the distribution of the responsible variable
a linear predictor of the covariates

Unit 4

PRACTICE QUESTION

a link function between the response variable and the linear predictor

(ii) A saturated model has as many parameters as there are data points and is therefore a perfect fit to the data.
It is not useful from a predictive point of view which is why it is not used in practice.
It is, however, a useful benchmark against which to compare the fit of other models.

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.

Ans:

(i) The saturated model is one where the number of parameters is the same as the data points,
i.e. the fitted values are the same as the fitted data.

(ii) The scaled deviance is twice the difference between the log likelihood values between the model in consideration and the saturated model.

(iii) (a) Pearson residuals are $\frac{y-\hat{\mu}}{\sqrt{var\hat{\mu}}}$ *where* $\hat{\mu}$ is the fitted response estimator.
The deviance residuals are sign$(y\text{-}\hat{\mu})d_i$ where d_i is the contribution of the i-th to the total deviances, i.e. $\sum_i^n d_i^2$ is the scaled deviance.
(b) The Pearson residuals tend to be skewed in non normal data while the deviance residuals tend to be symmetric and hence the normal assumption is more appropriate. For that reason the latter is preferred in actuarial applications.
(c) In the normal data, normal residuals these are identical.

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.

Unit 4

PRACTICE QUESTION

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

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

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

Ans:

(i) $$f(x)=\exp\left[\frac{x\theta-b(\theta)}{a(\varphi)}+c(x,\varphi)\right]$$

E(X) = $b'(\theta)$

V(X)= $a(\phi)b''(\theta)$

(ii) $a(\phi)2\ b'''(\theta)$

(iii) $\phi = \alpha$, $\theta = -1/\mu$, $a(\phi) = 1/\ \alpha$, $b(\theta) = -\log(-\theta)$, $c(x, \phi) = (\phi - 1)\log x + \phi \log \phi - \log\Gamma(\phi)$

(iv) $E(X - E(X))^3 = 2\mu^3/\ \alpha^2$

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.

Ans:

(i) $\widehat{\mu_1} = 2.4$, $\widehat{\mu_2} = 4.4$, $\widehat{\mu_3} = 5.6$

Unit 4

PRACTICE QUESTION

(ii) $df = 2$, test stat = 6.7103, Critical value = 5.991,
This value is above 5.991 which is the critical value at the upper 5% level and therefore conclude that mean claim rates are different.

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).

Ans:

(i) $\theta = log\ (\frac{\mu}{1-\mu})$, φ=n, a(φ) =1 / φ, b (θ)= log (1 + e $^{\theta}$), c(y, φ) = $log\ (\frac{\varphi}{y\varphi})$

(ii) E(y) = μ, V(y) = $\frac{\mu(1-\mu)}{n}$

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).

Ans:

(i) –

(ii) –

(iii) A GLM consists of three components:
a distribution for the data (Poisson, exponential, gamma, normal or binomial)
a linear predictor (a function of the covariates that is linear in the parameters)
a link function (that links the mean of the response variable to the linear predictor).

Unit 4

PRACTICE QUESTION

13. CS1A September 2019 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:

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

where $\underline{y_i}$ is the average number of claims per month for policy i:

$$\underline{y_i} = \sum_{j=1}^{m}\frac{Y_{ij}}{m}$$

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.

Ans:

(i) $\beta_i = (\underline{Y_i})$

(ii)

(iii) 9.957

Unit 4

PRACTICE QUESTION

14. CS1A September 2020 Question 7

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

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

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

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

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

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

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

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

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

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.

Ans:

Unit 4

PRACTICE QUESTION

(i) Option – A2
(ii) Option – A4
(iii) The expectation of X is correct. This is obtained by taking the derivative of b(theta). The standard deviation is not correct. In fact it is the variance that is s squared. It is obtained by taking the second derivative of b(theta) and multiply by a(phi).
(iv) A factor takes a categorical value and for a factor with k levels, there are generally k parameters. For a numerical variable, the value is included as such in the linear predictor and there is a single parameter in the model for each numerical variable.

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

Unit 4

PRACTICE QUESTION

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

Ans:

(i) A distribution of the response variable Y
A "linear predictor" η
A "link function" g

(ii) The distribution of the response D_x is a Poisson distribution. The linear predictor $\eta_x = a + b_x$. The link function is the logarithm since $log(E[D_x]) = \eta_x$.

(iii) Option – A1

(iv) Option - A3

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.

(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.

Ans:

(i) log $(\mu)=\alpha_i+\beta_i X_1$; where i=1,2,3 for low, medium and high pollutant respectively
$\mu=E(Y)$

(ii) α_i, i=1,2,3 are the coefficients of the main effect for pollutant concentration. We may also need the interaction term $\beta_i X_1$ if the effect of temperature on number of hospitalisations is different for each level of pollutant concentration

(iii) $\log(\mu) = -0.372 + 0.09 \times 19 + 0.298 - 0.076 \times 19$

(iv) These are not listed as X_2Low is used as the reference category or, equivalently, their effect is included in the intercept estimate

(v) Medium concentration has no significant effect, as compared to low concentration, while high concentration has a significant increasing effect for the number of hospital admissions.

17. CS1A April 2022 Question 7

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

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

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.

Unit 4

PRACTICE QUESTION

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: $cs = 7.52 + 0.06 \times opt$

Model 2: $cs = 3.6 + 0.04 \times opt + 2.32 \times l_r$

	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.

Ans:

(i) –

Unit 4

PRACTICE QUESTION

(ii) The gamma distribution is more suitable because claim sizes are always positive and their distribution is usually non-symmetrical.
(iii) We conclude that the covariate operational time is significant.
(iv) The variable "legal representation" is a factor that takes a categorical value (yes/no), while the operational time is a continuous covariate (or a variable taking a numerical value).
(v) The deviance improved significantly when using the legal representation as a second covariate. Therefore, legal representation is a significant covariate of the claim sizes and Model 2 is preferred.

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.

Ans:

(i) The linear predictor is

(ii) $a_i + beta_j + gamma_k + delta*x$

(iii) a_i is the gender effect on claim amount, for i = 1 – male, 2 – female
$beta_j$ is the experience effect, for j = 1,2,3,4
$gamma_k$ is the car group effect, for k = 1, …, 20
delta is the coefficient for the effect of numerical variable age (x)

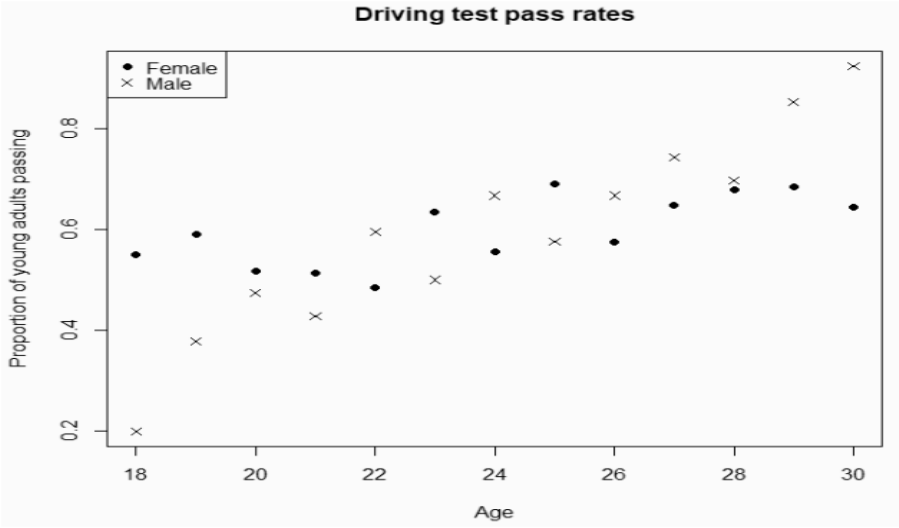
(iv) The linear predictor changes to

(v) $a_i + beta_j + gamma_k + delta_k*x$
where
a_i, $beta_j$, $gamma_k$ are as before,
$delta_k$ is the effect of age on claim amount, for different car groups, k = 1, …, 20

(vi) df = 19, deviance reduction= 112.2, Critical value = 36.19
We have strong evidence against H_0 and conclude that the model with the interaction term improves the fit significantly and should be preferred

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.

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:

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

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

Ans:

Unit 4

PRACTICE QUESTION

(i) The proportion passing appears to increase somewhat with age for males, and to a lesser extent for females.

(ii) $\eta = log\left(\frac{\mu}{1-\mu}\right) = \alpha_{\text{gender}} + \beta_{\text{gender}} \times \text{Age}$

Where gender is either male or female, $\mu = E(Y/n)$.

(iii) 56.6%

(iv) This expression means that for each additional year of age for a male, the ratio of the probability of passing against failing increases by a factor of $e^{0.209} = 1.232$

21. CS1A April 2023 Question 8

A space rocket contains six identical mechanical components that work independently of each other and need to be in operation for a successful launch. Data from simulated launches are available to establish the relationship between the number of damaged components (Y) on the rocket and air temperature (X, in degrees Fahrenheit). It is suggested to analyse the simulated data using a binomial Generalised Linear Model (GLM) with the canonical link function, where $Y \sim$ Binomial (6, p), p is the probability of a component being damaged, and the linear predictor has the form:

$$\beta_0 + \beta_1 X.$$

The analysis of the simulated data gave the following estimates for the model:

	Estimate	***Standard error***
β_0	11.6630	3.2963
β_1	−0.2162	0.0532

(i) Determine whether air temperature significantly affects the number of damaged components on the rocket by computing a suitable p-value.

(ii) Estimate (to two decimal places) the probability that a component will be damaged when the air temperature is 31 degrees Fahrenheit.

(iii) Estimate the expected value of the number of components that will be damaged when the air temperature is 31 degrees Fahrenheit.

It is believed that the launch is safe when at least five of these six components are not damaged.

(iv) (a) Calculate the probability that the launch is safe when the air temperature is 31 degrees Fahrenheit.

Unit 4

PRACTICE QUESTION

(b) Comment on the safety of the launch when the air temperature is 31 degrees Fahrenheit.

A second approach for analysing the simulated data was suggested, where a logarithmic link function would be used with the same GLM as used before.

(v) Comment on the suitability of the second approach.

Ans:

(i) Test stat = - 4.06, p-value = 0.00004
We have very strong evidence against H0 and conclude that air temperature significantly affects the number of damaged components.

(ii) 0.99

(iii) 5.94.

(iv) (a) $5.95e^{-10}$
(b) Launch is not safe when the air temperature is as cold as 31 degrees Fahrenheit

(v) This approach is not suitable as the analysis may give probability estimates that are greater than 1.

22. CS1A September 2023 Question 3

The probability mass function of a random variable, Y, is defined as:

$$P(Y = y) = \binom{k + y - 1}{y}(1 - p)^y p^k, \qquad y = 0, 1, 2, 3, \ldots$$

where k is a positive integer.

(i) State the distribution of the random variable Y. Your answer should include an explanation of the meaning of the distribution's parameters.

(ii) Identify which one of the following options gives the natural parameter θ, the scale parameter φ, and the relevant functions $b(\theta)$, $a(\varphi)$ and $c(y, \varphi)$ of the exponential family for this distribution.

Unit 4

PRACTICE QUESTION

A $\theta = \log(1-p),\ \varphi = 1,\ b(\theta) = -k\log p,\ a(\varphi) = 1,$
$c(y,\varphi) = \log\binom{k+y-1}{y}$

B $\theta = \log(p),\ \varphi = 1,\ b(\theta) = -k\log(1-p),\ a(\varphi) = 1,$
$c(y,\varphi) = \log\binom{k+y-1}{y}$

C $\theta = \log(1-p),\ \varphi = k,\ b(\theta) = -\log p,\ a(\varphi) = k,$
$c(y,\varphi) = \log\binom{k+y-1}{y}$

D $\theta = k\log(1-p),\ \varphi = 1,\ b(\theta) = k\log p,\ a(\varphi) = 1,$
$c(y,\varphi) = \log\binom{k+y-1}{k}.$

(iii) Derive, using the properties of the exponential family, the mean of Y.

(iv) Identify which **one** of the following options gives the variance of Y:

A $V[Y] = -\frac{1}{k}\frac{1-p}{p^2}$

B $V[Y] = \frac{1}{k}\frac{1-p}{p^2}$

C $V[Y] = k\frac{1-p}{p^2}$

D $V[Y] = k\frac{p}{(1-p)^2}.$

Ans:

(i) The random variable YY follows a negative binomial distribution and represents the number of failures before the kth success in trials occurring independently with probability of success $p\in(0,1)$

(ii) Option A

Unit 4

PRACTICE QUESTION

(iii) $E[Y] = k\frac{1-p}{p}$
(iv) Option C

22. CS1A April 2024 Question 2

The preparation time, in minutes, for coffees in a popular coffee shop has the following density:

$$f(y) = \frac{8}{m^4} y e^{-\frac{4y}{m}}, \qquad y \geq 0, m > 0$$

(i) Show that this density can be written in the form of the exponential family, also determining $b(\theta)$, $a(\varphi)$ and $c(y, \varphi)$

Two models, A and B, are proposed for the parameter m, the mean preparation time, for different types of coffee.

Model A: $\frac{1}{m} = u_i \quad i = 1, 2, 3, 4$

Model B: $\frac{1}{m} = \begin{cases} u & i = 1 \\ u + v & i = 2, 3, 4 \end{cases}$

where $i = 1, 2, 3, 4$ correspond to different types of coffee.

(ii) Comment on the mean preparation times for different types of coffee, proposed by the two models.

Ans:

(i) $\theta = -1/m$
(ii) Model A:
All four mean preparation times are different from each other.
Model B:
The mean preparation times for types $i = 2$, $i = 3$ and $i = 4$ are the same. On the other hand, the mean preparation time for type $i = 1$ could potentially be different (due to the absence of the v term).

Unit 4

PRACTICE QUESTION