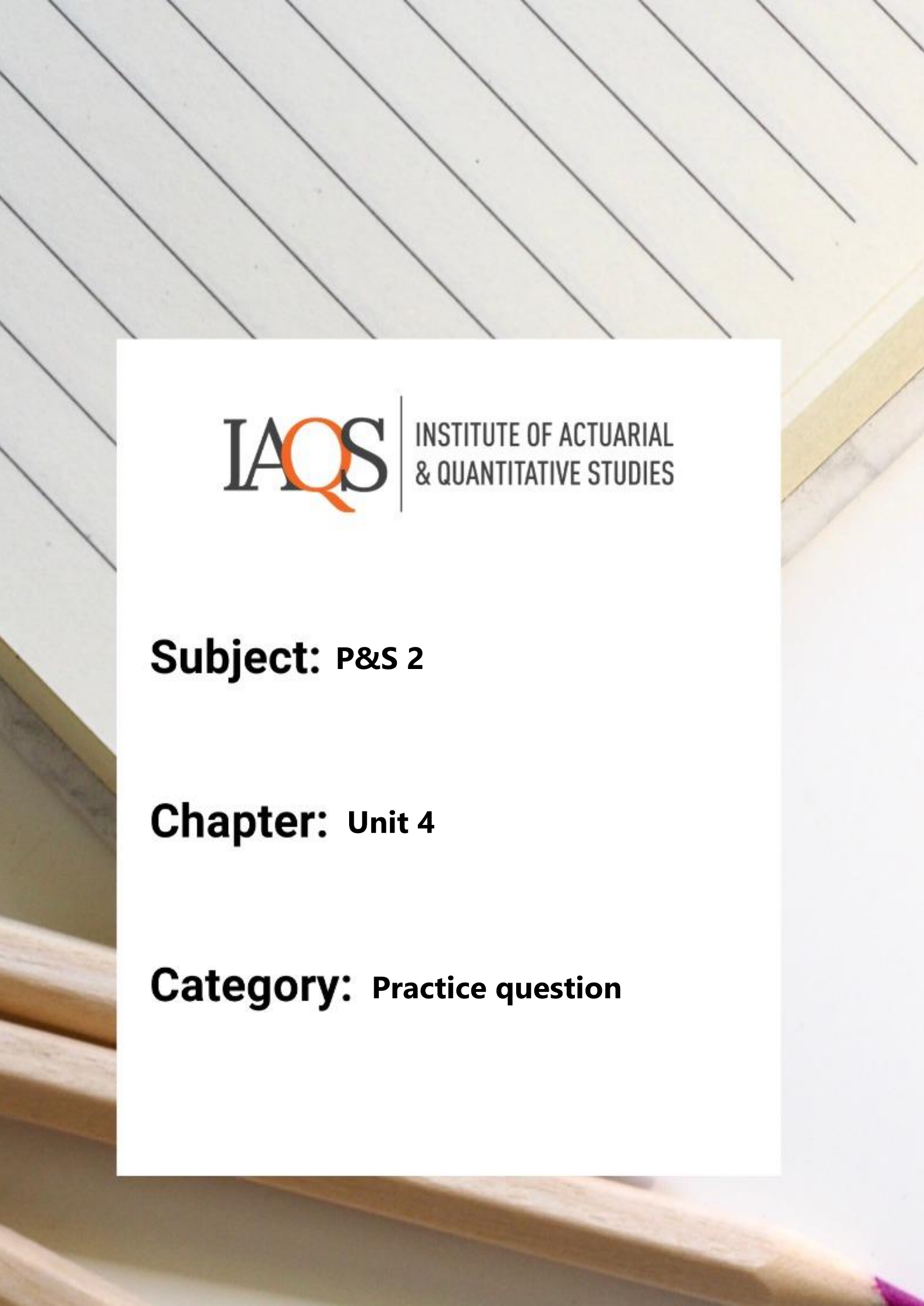


INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: P&S 2

Chapter: Unit 4

Category: Practice question

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 October 2012 Question 10

The number of hours that people watch television per day is the subject of .an empirical studythat is carried out in four regions in a country. Five people are randomly selected in each of theregions and are asked about the average number of hours per day that they spent watching television during the last year. The results are shown in the following table, with the last column showing the average in each region.

Region 1	2.0	1.1	0.2	3.8	2.8	1.98
Region 2	1.2	1.0	0.9	1.1	1.6	1.16
Region 3	2.5	2.0	2.6	2.4	2.3	2.36
Region 4	1.2	1.7	1.0	1.8	1.3	1.40

Based on the above observations the following ANOVA table was obtained:

Source of variation	d.f	SS	MSS
Between regions	...	4,4655	...
Residual	...	8.892	...

(i) State the mathematical model underlying the. one-way analysis of variance togetherwith all associated assumptions.

(ii) Complete the ANOVA table.

3. CT6 April 2013 Question 8

A random sample of 10 independent claim amounts was taken from each of three different regions and an analysis of variance was performed to "compare the mean level of claims in these regions. The resulting ANOVA given below.

Unit 4

PRACTICE QUESTION

Source	d.f.	SS	MSS
Between regions	2	4.439.7	2.219.9
Residual	27	10.713.5	396.8
Total	29	15,153.2	

(i) Perform the appropriate, F test to determine whether there are significant differences between, the mean claim amounts for the three regions. You should state clearly the hypotheses of the test.

The three sample means were:

Region	A	B	C
Sample mean	147.47	154.56	125.95

It was of particular interest to compare regions A and B.

(ii) (a) Calculate a 95% confidence interval for the difference between the means forregions A and B.

(b) Comment on your answer in part (ii)(a) given the result of the F test performed inpart (I).

4. 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:

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

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

Unit 4

PRACTICE QUESTION

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.

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 April 2015 Question 6

Annual numbers of claims on three different types of insurance policy follow a Poisson distribution with parameter μ_i for 1, 2, 3 i = 1,2,3… Data for the last four years is given in the table below.

	Year				
Type	*1*	*2*	*3*	*4*	*Total*
1	5	5	0	1	11
2	2	5	4	5	16
3	5	6	4	5	20

(i) Derive the maximum likelihood estimate of m1 and calculate the corresponding estimates of μ_2 and μ_3

(ii) Test the hypothesis that μ_1, μ_2 and μ_3 are equal using the scaled deviance.

Unit 4

PRACTICE QUESTION

8. 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:

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

(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.

Unit 4

PRACTICE QUESTION

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).

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:

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

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

Unit 4

PRACTICE QUESTION

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

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:

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

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

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

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

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

(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 $\mathrm{SD}(X) = s^2$. In your answer you may denote θ by theta and ϕ by phi.

Unit 4

PRACTICE QUESTION

(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.

(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:

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

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

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

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

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: $cs = 7.52 + 0.06 \times opt$

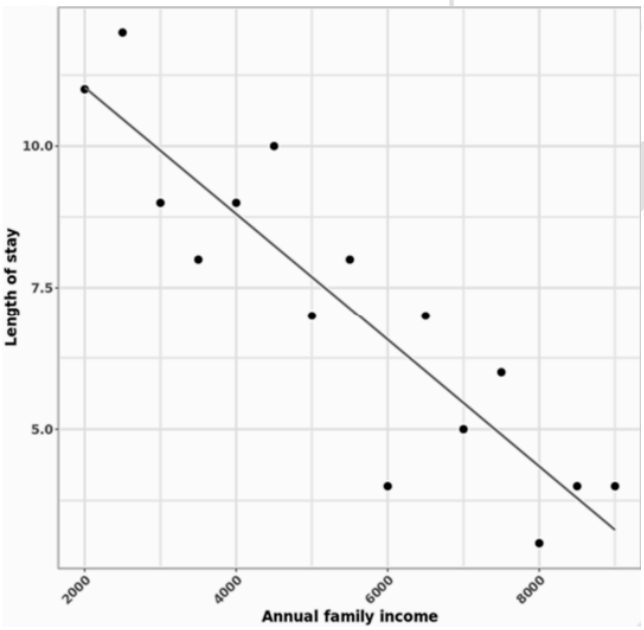
Model 2: $cs = 3.6 + 0.04 \times opt + 2.32 \times l_r$

	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.

18. CS1A April 2022 Question 10

A random sample of the records of a certain hospital yielded the following information on the length of hospital stay in days (l_i) and the annual family income (ai, rounded to the nearest £500) of 15 discharged patients. An analyst believes that the relationship between these two variables is linear. The graph below depicts the scatter plot of the annual family income against the length of stay and the simple linear regression line fitted by the analyst.



Summary statistics for these data are given below:

$$\sum a_i = 82{,}500,\ \ \sum a_i^2 = 523{,}750{,}000,\ \ \sum a_i l_i = 510{,}500,\ \ \sum l_i = 107,\ \ \sum l_i^2 = 871.$$

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

(ii) Determine the equation of the simple regression line.

(iii) Perform an ANOVA test to determine whether the slope of the regression line is significantly different from zero.

(iv) Calculate Pearson's correlation coefficient between the annual family income and the length of hospital stay.

(v) Perform a statistical test to determine whether Pearson's correlation coefficient for the corresponding population is significantly different from −0.8.

(vi) Identify which one of the following options gives an approximate 95% confidence interval for Pearson's correlation coefficient for the corresponding population:

A (–2.027, –0.896)

B (–0.966, –0.714)

C (–0.989, –0.683)

D (–0.908, –0.794)

19. 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.

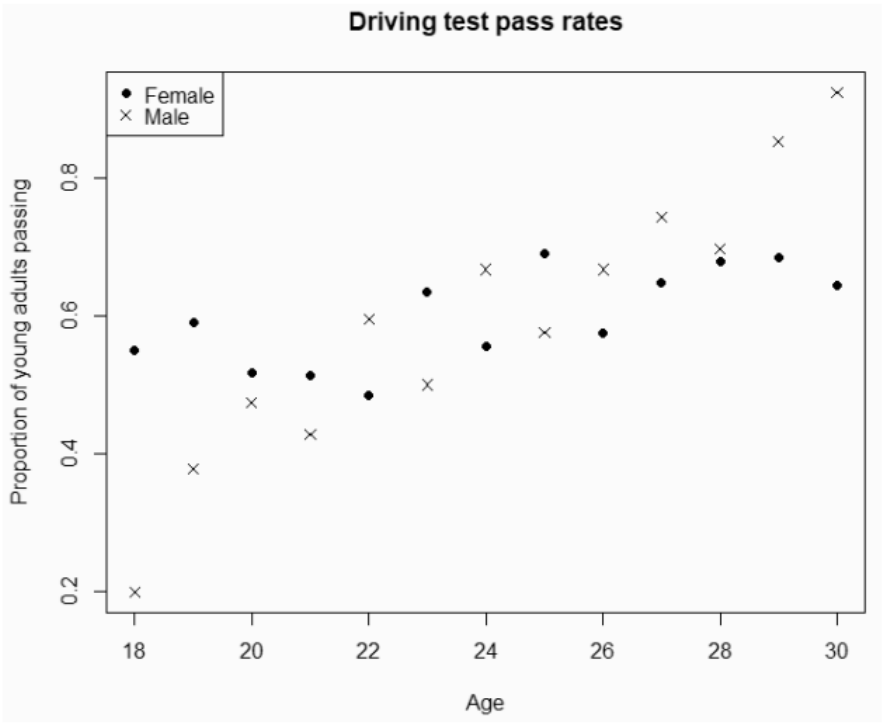
You should explain all the terms used in the new linear predictor.

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.

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:

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

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