11.3 The sums of the squares of the errors in a regression analysis are found to be:

$$SS_{REG} = \sum(\hat{y}_i - \bar{y})^2 = 6.4 \qquad SS_{RES} = \sum(y_i - \hat{y}_i)^2 = 3.6 \qquad SS_{TOT} = \sum(y_i - \bar{y})^2 = 10.0$$

Calculate the coefficient of determination and explain what this represents.

11.6 The share price, in pence, of a certain company is monitored over an 8-year period. The results are shown in the table below:

Exam style

Time (years)	0	1	2	3	4	5	6	7	8
Price	100	131	183	247	330	454	601	819	1,095

$$\sum(x_i - \bar{x})^2 = 60 \qquad \sum(y_i - \bar{y})^2 = 925{,}262 \qquad \sum(x_i - \bar{x})(y_i - \bar{y}) = 7{,}087$$

An actuary fits the following simple linear regression model to the data:

$$y_i = \alpha + \beta x_i + e_i \qquad i = 0,1,\ldots,8$$

where $\{e_i\}$ are independent normal random variables with mean zero and variance σ^2.

(i) Determine the fitted regression line in which the price is modelled as the response and the time as an explanatory variable. [2]

(ii) Calculate a 99% confidence interval for:

(a) β, the true underlying slope parameter

(b) σ^2, the true underlying error variance. [5]

(iii) (a) State the 'total sum of squares' and calculate its partition into the 'regression sum of squares' and the 'residual sum of squares'.

(b) Use the values in part (iii)(a) to calculate the 'proportion of variability explained by the model' and comment on the result. [5]

(iv) The actuary decides to check the fit of the model by calculating the residuals.

(a) Complete the table of residuals (rounding to the nearest integer):

Time (years)	0	1	2	3	4	5	6	7	8
Residual	132		−21	−75		−104	−75	25	

(b) Use a dotplot of the residuals to comment on the assumption of normality.

(c) Plot the residuals against time and hence comment on the appropriateness of the linear model. [7]

11.7 Exam style

A schoolteacher is investigating the claim that class size does not affect GCSE results. His observations of nine GCSE classes are as follows:

Class	X1	X2	X3	X4	Y1	Y2	Y3	Y4	Y5
Students in class (c)	35	32	27	21	34	30	28	24	7
Average GCSE point score for class (p)	5.9	4.1	2.4	1.7	6.3	5.3	3.5	2.6	1.6

$$\sum c = 238 \quad \sum c^2 = 6,884 \quad \sum p = 33.4 \quad \sum p^2 = 149.62 \quad \sum cp = 983$$

(i) Determine the fitted regression line for p on c. [3]

(ii) Class X5 was not included in the results above and contains 15 students. Calculate an estimate of the average GCSE point score for this individual class and specify the standard error for this estimate assuming the full normal model. [4]

11.10 Exam style

The government of a country suffering from hyperinflation has sponsored an economist to monitor the price of a 'basket' of items in the population's staple diet over a one-year period. As part of his study, the economist selected six days during the year and on each of these days visited a single nightclub, where he recorded the price of a pint of lager. His report showed the following prices:

Day (i)	8	29	57	92	141	148
Price (P_i)	15	17	22	51	88	95
$\ln P_i$	2.7081	2.8332	3.0910	3.9318	4.4773	4.5539

$$\sum i = 475 \quad \sum i^2 = 54,403 \quad \sum \ln P_i = 21.5953 \quad \sum (\ln P_i)^2 = 81.1584 \quad \sum i \ln P_i = 1,947.020$$

The economist believes that the price of a pint of lager in a given bar on day i can be modelled by:

$$\ln P_i = a + bi + e_i$$

where a and b are constants and the e_i's are uncorrelated $N(0,\sigma^2)$ random variables.

(i) Estimate a, b and σ^2. [5]

(ii) Calculate the linear correlation coefficient r. [1]

(iii) Calculate a 99% confidence interval for b. [2]

(iv) Determine a 95% confidence interval for the average price of a pint of lager on day 365:

(a) in the country as a whole

(b) in a randomly selected bar. [7]

[Total 15]

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. [2]

(ii) Determine the correlation coefficient between Y and T. [4]

(iii) Perform a statistical test using Fisher's transformation to determine whether the population correlation coefficient is significantly different from zero. [6]

(iv) Determine the parameters of a linear regression, including writing down the equation. [3]

[Total 15]

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

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

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

(i) Determine the linear regression equation of d on g. [5]

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. [4]

(iii) Determine a 95% confidence interval for the mean estimate of d on a beach with a slope of exactly 3 degrees. [5]

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

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

[4]

[Total 18]