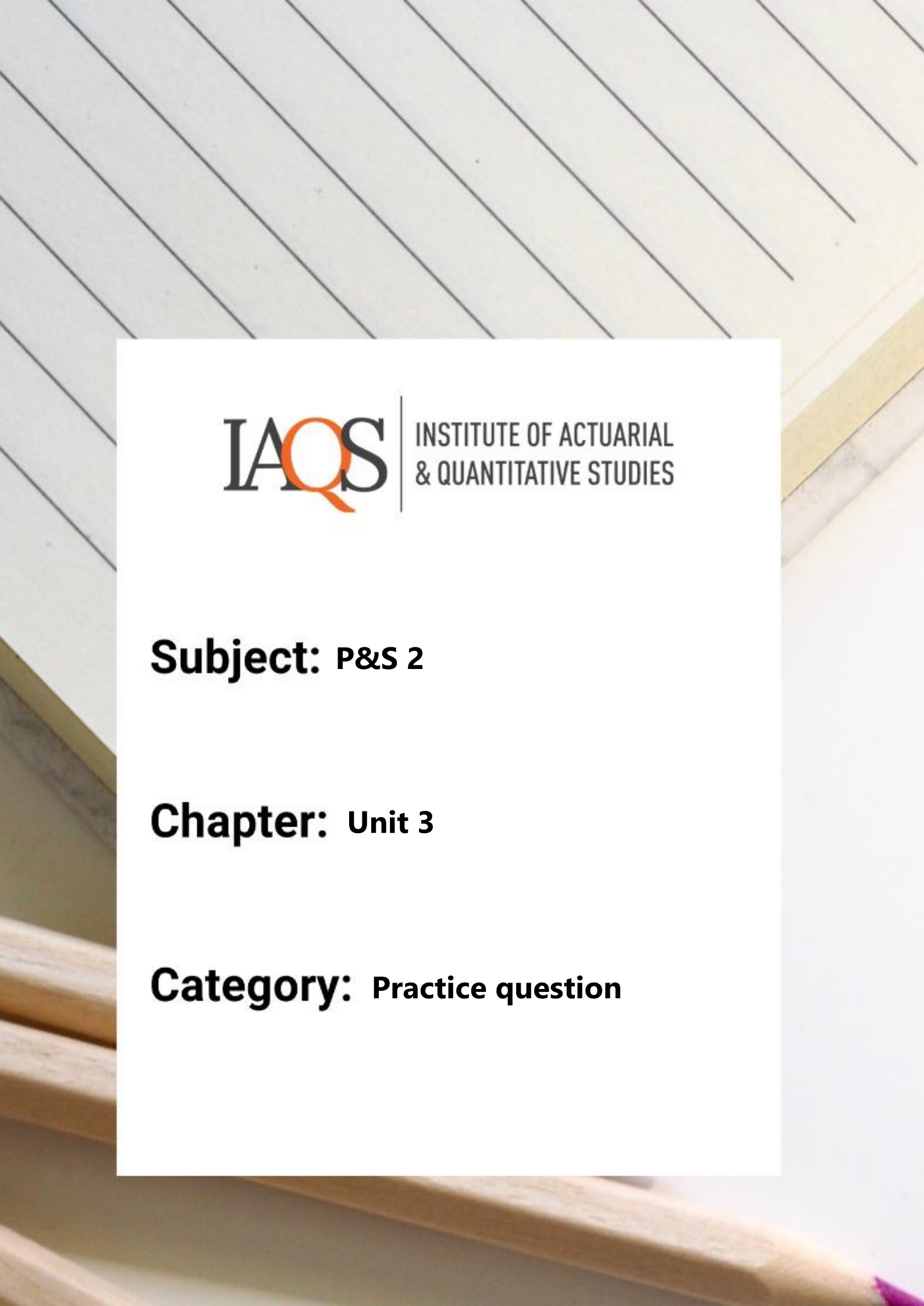


INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

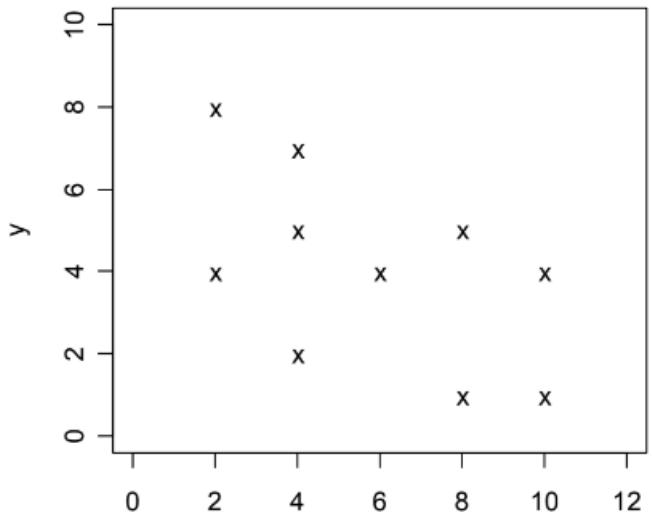
**Subject:** P&S 2

**Chapter:** Unit 3

**Category:** Practice question

**1. CT6 October 2011 Question 10**

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




For these data: $\sum x = 58, \sum x^2 = 420, \sum y = 41, \sum y^2 = 217, \sum xy = 202$

(i) (a) Calculate the value of the coefficient of determination ($R^2$) for the data.

(b) Determine the equation of the fitted least-squares line of regression of y on x .

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

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

(b) Calculate the standard error of this estimate.

Suppose the observation (x = 10, y = 8) is added to the existing data. The coefficient of determination is now $R^2 = 0.07$.

(iv) Comment briefly on the effect of the new observation on the fit of the linear model.

**2. CT6 April 2012 Question 13**

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

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

The last column contains the sum of all eight columns.

From these values we obtain the following results:

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

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

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

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

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

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

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

**3. CT6 October 2012 Question 13**

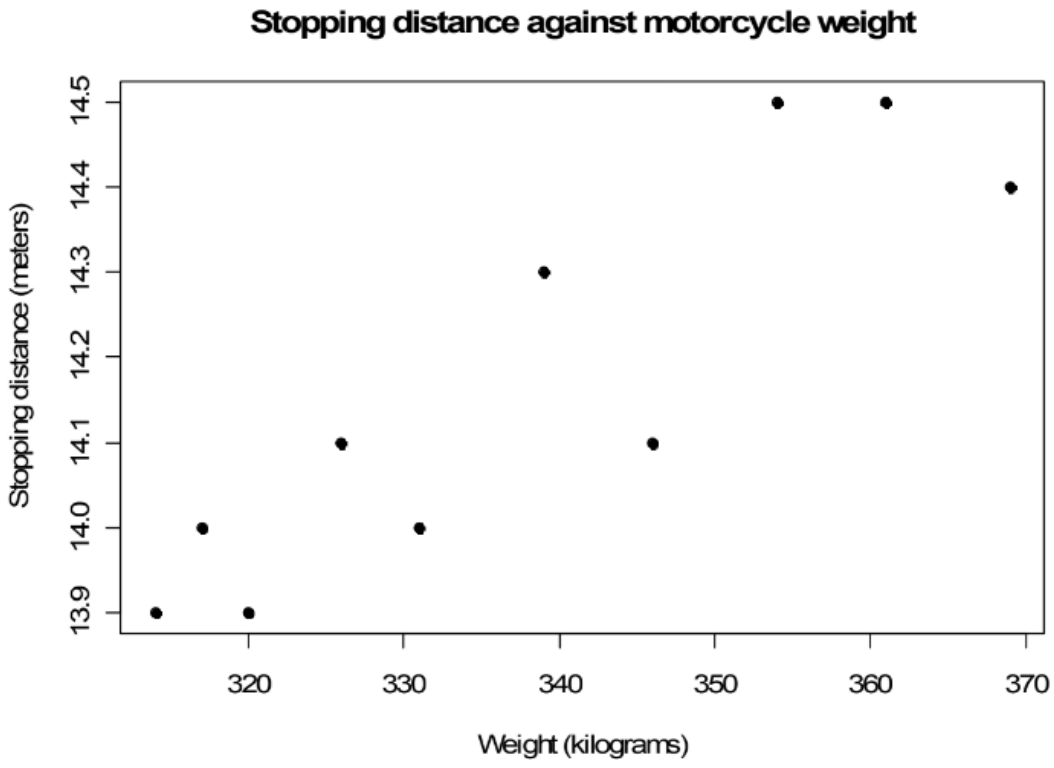
The following data give the weight, in kilograms, of a random sample of 10 different models of similar motorcycles and the distance, in metres, required to stop from a speed of 20 miles per hour.

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

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

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

A scatter plot of the data is shown below.




(i) (a) Comment briefly on the association between weight and stopping distance, based on the scatter plot.
(b) Calculate the correlation coefficient between the two variables.

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

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

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

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

**3. CT6 October 2013 Question 10**

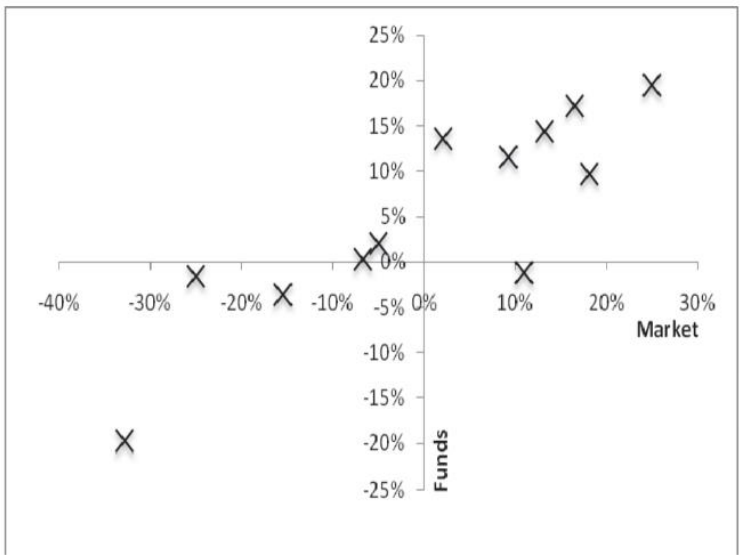
An analyst wishes to compare the results from investing in a certain category of hedge funds, f, with those from the stock market, x. She uses an appropriate index for each, which over 12 years each produced the following returns (in percentages to one decimal place).

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

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

It is assumed that observations from different years are independent of each other.

Below is a scatter plot of market returns against fund returns for each year.




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

**Unit 3**

**PRACTICE QUESTION**

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

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

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

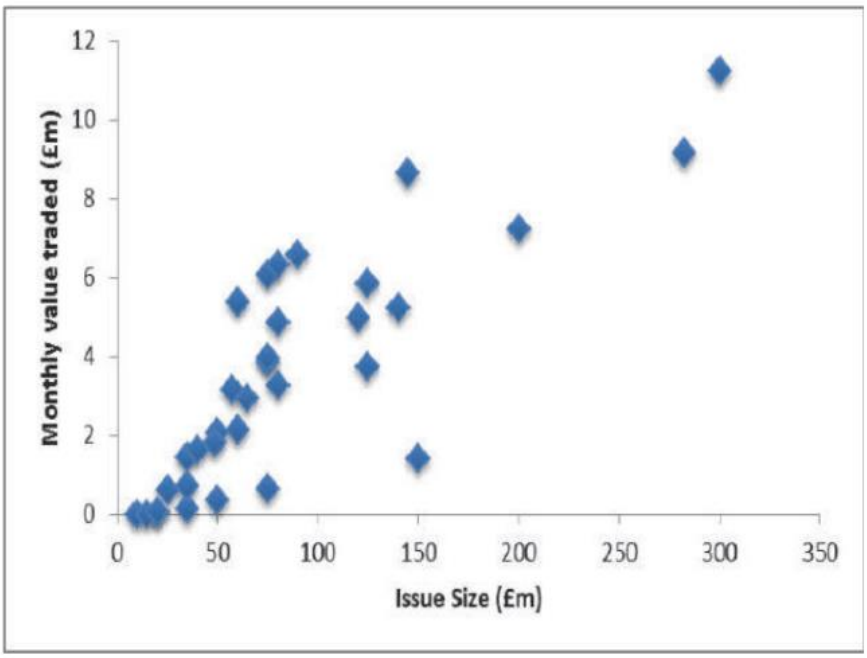
(iii) Calculate the parameters for a linear regression of the fund index on the market index.

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

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

**4. CT6 April 2014 Question 10**

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




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

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

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

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

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

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

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

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

(v) Comment on the colleague's observation.

**5. CT6 September 2014 Question 10**

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

| Region, I | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Average |
|---|---|---|---|---|---|---|---|---|---|
| Alcohol units per week, $x_i$ | 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.

**Unit 3**

**PRACTICE QUESTION**

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

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

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

**6. CT6 October 2015 Question 5**

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

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

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

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

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

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

**7. CT6 October 2015 Question 11**

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

**8. CT6 April 2016 Question 11**

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

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

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

**Unit 3**

**PRACTICE QUESTION**

(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 ($\varepsilon$) independently follow a $N(0, \sigma^2)$ distribution, with $\sigma^2$ s being an unknown parameter.
(ii) Determine the fitted line of the regression model.

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

The article suggests that there is a 'clear relationship' between mileage and selling price of the car.
(b) Comment on this suggestion based on the confidence interval obtained in part (iii)(a).

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

**9. CT6 April 2017 Question 10**

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

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

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

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

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

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

(iv) (a) Plot the data from the table above.
(b) Comment on the plot suggesting what the geologist might do to improve her analysis.

**Unit 3**

**PRACTICE QUESTION**

**10. CT6 September 2017 Question 10**

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

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

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

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

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

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

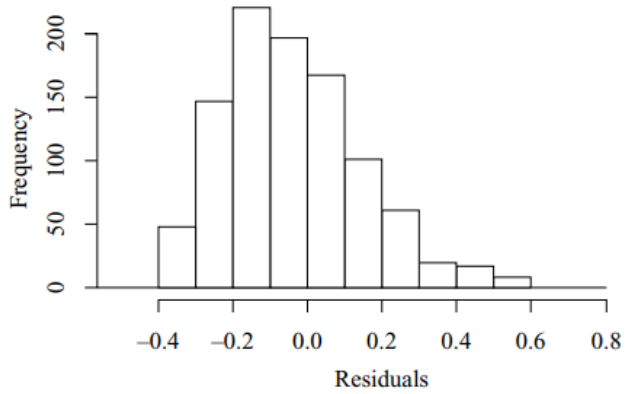
(iii) Perform a statistical test using Fisher's transformation to determine whether the population correlation coefficient is significantly different from zero.

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

**11. CS1A September 2019 Question 6**

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

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




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

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

for which

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

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

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, $\sigma^2$) distribution, and where a, b and s2 are unknown parameters.

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

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

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

**Unit 3**

**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 $\mu_X$ and $\mu_Y$,and variances $\sigma_X^2$ and $\sigma_Y^2$. $S_X$ 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.

A1 $y_i = a + bx_i$

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

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

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

**Unit 3**

**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.

**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 $\sigma^2$. In a sample of size n = 10, the following statistics have been observed:

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

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

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

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

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

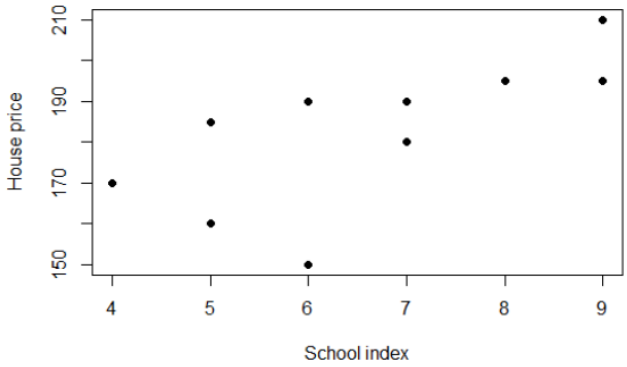
**14. CS1A September 2020 Question 10**

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

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

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

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




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

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

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

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

A1 2.295

A2 6.071

A3 2.743

A4 4.009

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

**Unit 3**

**PRACTICE QUESTION**

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

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

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

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

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

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

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

(b) Calculate the value of the test statistic.

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

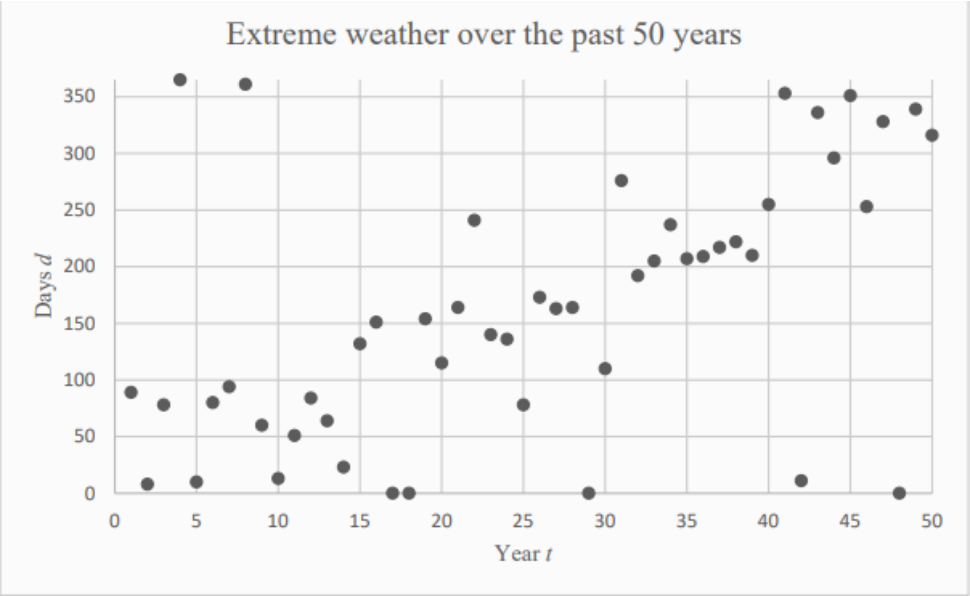
(d) Write down the conclusion of the test at the 5% level of significance, including the relevant critical value(s) from the Actuarial Formulae and Tables.

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

**15. CS1A April 2021 Question 8**

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

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




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

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

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

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

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

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

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

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

**Unit 3**

**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).

**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 $\hat{b} = -5.506 \times 10^{-5}$

**Unit 3**

**PRACTICE QUESTION**

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).

**Unit 3**

**PRACTICE QUESTION**

**17. CS1A April 2022 Question 9.**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

| Model | $R^2$ | Adjusted $R^2$ |
|---|---|---|
| $X_1$ | 0.7322 | 0.7167 |
| $X_1 + X_4$ | 0.8018 | 0.7712 |

**Unit 3**

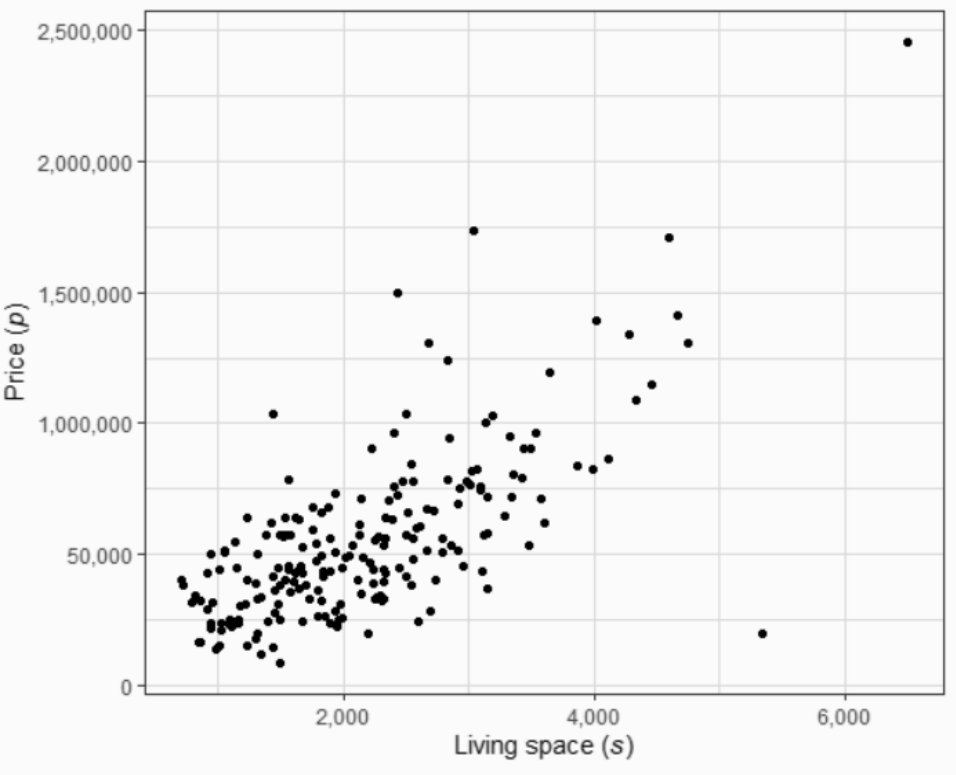
**PRACTICE QUESTION**

| | | |
|---|---|---|
| $X_1 + X_4 + X_3$ | 0.8253 | 0.7805 |
| $X_1 + X_4 + X_3 + X_2$ | 0.8259 | 0.7684 |

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

**18. CS1A September 2022 Question 9**

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




(i) Comment on the graph.

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

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

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

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

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

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

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

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

(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).