Tutorial 11

CB2200 Business Statistics, YANG Yihang

November 24, 2020

Topic 7-Question 8

Q8

a) Grant, Inc., a manufacturer of women's dress blouses, knows that its brand is carried in 19 percent of the women's clothing stores in Hong Kong. Grant recently sampled 85 women's clothing stores in mainland China and found that 14 percent of the stores carried the brand.

(i) At the 0.05 level of significance, is there evidence that Grant has poorer distribution in mainland China than it does in Hong Kong?
(ii) Interpret the decision you made in (i) in the situation being examined.
(iii) What is the p-value in (i)? Do you make the same decision as in (i) at $\alpha = 0.05$ if you use the p-value approach?
(iv) Suppose that the sample size $n = 85$ is fixed and further suppose that the penalty of committing type II is serious, which α value (0.05 or 0.10) do you choose for the hypothesis testing? Briefly explain your choice.

b) From past records, a charity has found that 42% of donors in a year will denote again in the next year. A random sample of 300 donors from last year was taken.

(i) What is the standard error of the sample proportion who will donate again this year?
(ii) What is the probability that the sample proportion is between 0.40 and 0.45?
(iii) Without doing the calculations, state in which of the following ranges the sample proportion is more likely to lie: 0.39 to 0.41, 0.41 to 0.43, 0.43 to 0.45.
(iv) Interpret the answer obtained in (ii).

Topic 7-Question 8

Q8

a)

(i) Let π be the population proportion of stores carried the brand

$H_0 : \pi \geq 0.19$

$H_1 : \pi < 0.19$

$\because n = 85 \geq 30$ np=11.9>5 n(1-p)=73.1>5

$\therefore$ Sampling distribution of p is approximately normal.

$\alpha = 0.05$ Critical Value $= -Z_\alpha = -Z_{0.05} = -1.645$

Reject H_0 if $Z < -1.645$

$$Z = \frac{p_s - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.14 - 0.19}{\sqrt{\frac{0.19 \times 0.81}{85}}} = -1.1751$$

$\because Z = -1.1751 > -1.645$

$\therefore$ We do not reject H_0. There is insufficient evidence that Grant has poorer distribution in Mainland China than it does in Hong Kong.

(ii) There is insufficient evidence that Grant has poorer distribution in Mainland China than it does in Hong Kong.

(iii) p-value= $\Pr(Z \leq -1.18) = 0.1190$

Reject H_0 if p-value <0.05

$\because$ p-value = 0.1190> 0.05

$\therefore$ We do not reject H_0 and making the same decision as in (i)

(iv) Choose $\alpha = 0.1$

For a fixed sample size, a larger value of α would correspond to a smaller value of β, that can decrease the penalty of committing type II error.

Topic 7-Question 8

b)

(i) Standard error of sample proportion $= \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.42 \times 0.58}{300}} = 0.0285$

(ii) $\because n = 300 \geq 30 \quad n\pi = 300 \times 0.42 = 126 \geq 5 \quad n(1-\pi) = 300 \times 0.58 = 174 \geq 5$

$\therefore$ Sampling distribution of p_s is approximately normal.

$$\Pr(0.4 < p_s < 0.45) = \Pr\left(\frac{0.4-0.42}{0.0285} < Z < \frac{0.45-0.42}{0.0285}\right) = \Pr(-0.70 < Z < 1.05)$$
$$= 0.8531 - 0.242 = 0.6111$$

(iii) The range of 0.41 to 0.43 is more likely to lie because this range contains the population proportion that is 0.42.

(iv) In the sample size 300, 61.11% of sample will be expected to have the sample proportions between 0.4 and 0.45.

Topic 7-Question 9

Q9

a) In order to estimate the unemployment rate of Hong Kong, a random sample of 8500 people was selected in 2002 and 618 people were found to be unemployed. Find a 95% confidence interval for the unemployment rate of Hong Kong in 2002. Give your answer to the fourth decimal place.

b) If you want to be 95% confident of estimating the unemployment rate of Hong Kong in part (a) to within ±0.2%, what sample size is needed?

c) According to the report given by the Census and Statistics Department of Hong Kong, the actual unemployment rate of Hong Kong in 2002 is 7.3%. In a survey of 620 people in Shatin, 34 people were found to be unemployed. Is there evidence that the Shatin unemployment rate was lower than the Hong Kong unemployment rate at the 0.05 level of significance?

Topic 7-Question 9

Q9

a) Let p be the proportion of unemployment rate of Hong Kong in 2002

$n = 8500 \geq 30$

$np = 8500 \times (\frac{618}{8500}) = 618 \geq 5$

$n(1-p) = 85000 \times \left(1 - \frac{618}{85000}\right) = 7882 \geq 5$

$\therefore$ sampling distribution p is normal

assume population follows binomial.

For 95% Confidence Interval,

$$= 0.0727 \pm 1.96\sqrt{\frac{0.0727(0.9273)}{8500}}$$

$= [0.0672, 0.0782]$

$\therefore$ We are 95% confident that the population proportion of Hong Kong unemployment rate is estimated to be between 0.0672 and 0.0782.

b) Sample size

$$n = \frac{z_{\frac{0.05}{2}}^2 p(1-p)}{E^2} = 64750$$

Topic 7-Question 9

c) Let π be the proportion of unemployment rate of Shatin in 2002

$H_0 : \pi \geq 0.073$

$H_1 : \pi < 0.073$

$\because n$ =620>30,

np = 34 >5

n(1-p) = 586 >5

$\therefore$ the sampling distribution of p is approximately normal

$$p = \frac{34}{620} = 0.0548$$

$$p \sim N(0.073, \sqrt{\frac{0.073(1-0.073)}{620}}^2)$$

Reject H_0 if $Z < -1.645$

$$Z = \frac{p-\pi}{\sqrt{\frac{\pi\ (1-\pi\ )}{n}}} = \frac{0.0548-0.073}{\sqrt{\frac{0.073(1-0.073)}{620}}} = -1.738 < -1.645$$

Therefore we reject H_0. There is sufficient evidence that the unemployment in Shatin is lower than Hong Kong in 2002.

Topic 7-Question 10

Q10

a) Suppose there is population with 4 customers. The interested categorical variable is the proportion of customers that prefer your brand. In this population, one of them prefers your brand, and the others do not.

(i) Find the population proportion (π) if customers that prefer your brand.

In the process of developing sampling distribution, all possible samples (taken with replacement) of size n = 2 are obtained. The sample proportion (p) of customers that prefer your brand is considered as the sample statistic.

(ii) Develop the probability distribution of the sample proportion (p).
(iii) Show that the sample proportion p is an unbiased estimator for π.
(iv) Does the sampling distribution of p follow a Normal Distribution? Explain.

b) Your company markets a computer medical diagnostic programme. The programme scans the results of medical test on white blood cells. The patient could be referred to a doctor if the proportion of white blood cells is at most 5%, and however the patient can leave if the proportion of white blood cells is more than 5%.

In studying the effectiveness of the programme, your null hypothesis is that the population proportion of white blood cells is at most 5%, with the alternative hypothesis being that the population proportion of white blood cells is more than 5%.

(i) Would you rather make a type I or a type II error? Explain.
(ii) You want to have 90% confidence of estimating the population proportion of white blood cells to within ±1.95%. Because you have not previously undertaken such a study, there is no information available from past data. Determine the sample size needed.

Topic 7-Question 10

Q10

a)

(i) $\pi = \frac{1}{4} = 0.25$

(ii) Possible value of sample proportion of preferring the brand:

	Not prefer	Not prefer	Not prefer	prefer
Not prefer	0	0	0	0.5
Not prefer	0	0	0	0.5
Not prefer	0	0	0	0.5
prefer	0.5	0.5	0.5	1

Probability distribution of the sample proportion:

p	0	0.5	1
Pr(p)	$\frac{9}{16} = 0.5625$	$\frac{6}{16} = 0.375$	$\frac{1}{16} = 0.0625$

Possible value of p from the above table: 0, 0.5, 1

(iii) $\mu p = 0*0.5625+0.5*0.375+1*0.0625=0.25$

From part (i), $\pi=0.25$

=> $\mu p = \pi$

=> sample proportion p is an unbiased estimator for π

(iv) $n=2 < 30$, $np=2*0.25=0.5<5$, $n(1-p)= 2*(1-0.25)=1.5<5$

=> sample distribution of p does not follow a normal distribution

Topic 7-Question 10

b)

(i) H_0: $\pi \leq 5\%$ vs H_1: $\pi > 5\%$

Type I error: let unhealthy patient with $\leq 5\%$ white blood cells leave, resulting in unhealthy patients are not treated

Type II error: refer healthy patient with $> 5\%$ white blood cells to doctor, resulting in more cost is incurred or more patients are sent to doctor

Comparing the above type I and II error, type I error is more serious. We would rather make a type II error.

(ii) If there is no information available from past data,

α=1-90%=0.10, $Z_{\frac{0.10}{2}} = 1.645$

E=1.95%=0.0195

$\pi \approx 0.5$

Sample size, $n=\frac{1.645^2 \times 0.5 \times (1-0.5)}{0.0195^2}=1779.11 \cong 1780$ (round up)

Topic 7-Question 11

Q11

MTR Corporation has to conduct surveys regularly to evaluate its service quality. According to previous studies, 87% of the passengers refuse to take part in such surveys.

a) At minimum, how many passengers must be sampled so that the 95% confidence interval will specify the population proportion of responded passengers to within ±6%?

MTR recently sampled 350 passengers, and only 28 of them responded to the survey.

b) At the 0.05 level of significance, is there evidence that the response rate has been dropped?

c) Find a 95% confidence interval for the population proportion of passengers who are willing to respond to the survey.

Topic 7-Question 11

Q11

X is the number of passengers responded to the survey

π is the population proportion of passenger responded to the survey

a) $$n = \frac{(1.96)^2(0.13)(1-0.13)}{(0.06)^2} = 120.69 = 121 \text{ (round-up)}$$

b) $H_0: \pi \geq 0.13$

$H_1: \pi < 0.13$

As $n = 350 > 30$; $np = 28 > 5$; $n(1-p) = 322 > 5$

→ $p \sim N$

→ use Z test

At $\alpha = 0.05$, reject H_0 if $Z < -1.645$

$$p = \frac{28}{350} = 0.08$$

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.08 - 0.13}{\sqrt{\frac{0.13(1-0.13)}{350}}} = -2.7815$$

As $Z = -2.7815 < -1.645$, reject H_0.

There is sufficient evidence that the response rate has been dropped.

Topic 7-Question 11

c) 95% CI for π

$$= p \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} = 0.08 \pm 1.96 \sqrt{\frac{0.08(1-0.08)}{350}} = 0.08 \pm 0.0284 = [0.0516, 0.1084]$$

We are 95% confident that the true unknown population proportion of passengers who are willing to response to the survey is between 0.0516 and 0.1084 (i.e. 5.16% or 10.84%).

Intended Learning Outcomes

After this tutorial, you may know how to

- interpret the coefficient of covariance,
- construct linear regression model.

Summary---Topic 8: Simple Linear Regression

Covariance and Correlation Coefficient (indicating linear relationship)

	Covariance (scale dependent)	**Correlation Coefficient $\in [-1, 1]$ (scale independent)**
	• > 0: positive linear relationship between X and Y • < 0: negative linear relationship between X and Y • = 0: no linear relationship between X and Y *(only implies X and Y are not linearly related, but does not imply X and Y are not related)*	
Population	$\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$	$\rho_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(X_i - \mu_X)^2 \sum_{i=1}^{N}(Y_i - \mu_Y)^2}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
Sample	$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$	$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{S_X S_Y}$

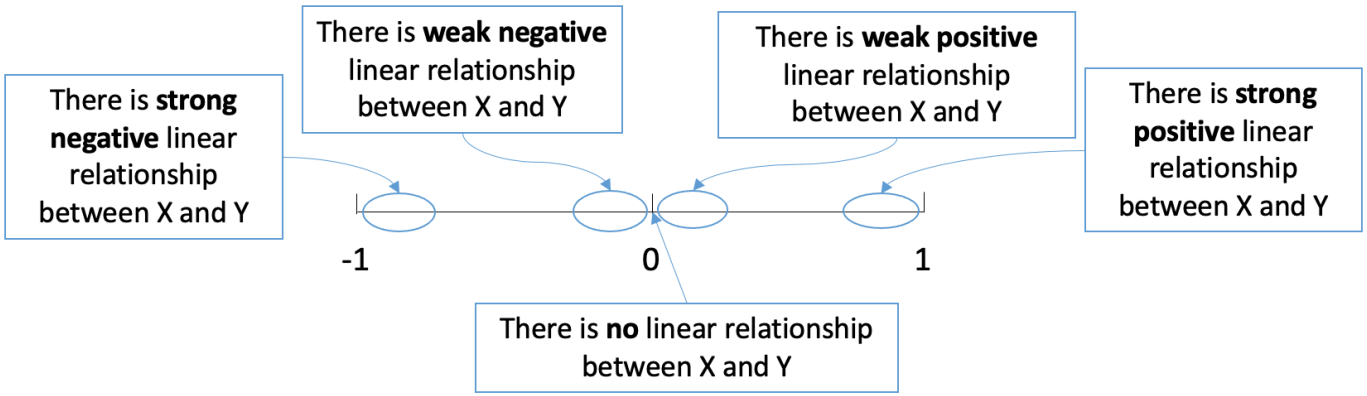
- The sign of ρ_{XY} (r_{XY}) is the same as that of σ_{XY} (S_{XY})

Correlation Coefficient

- Covariance does not have any upper boundary or lower boundary, while **correlation coefficient must range between -1 and 1**.

- The **magnitude of correlation coefficient** can measure the strength of the linear relationship:
  The closer the absolute value of correlation coefficient to 1 (i.e. correlation coefficient is close to 1 or -1), the stronger is the relationship between X and Y.

- Interpretation of correlation coefficient:
  There is (strong/weak) (positive/negative) linear relationship between X and Y

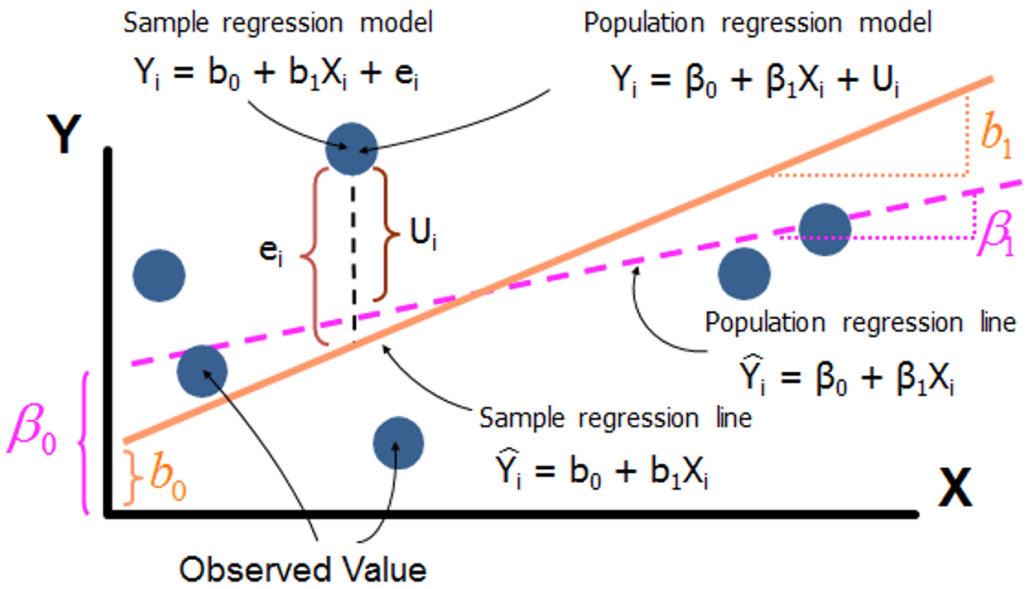


Simple Linear Regression

Relate the dependent variable with a single independent variable by a linear equation

Population simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$	**Sample simple linear regression model** $Y_i = b_0 + b_1 X_i + e_i$
Population regression line $\hat{Y}_i = \beta_0 + \beta_1 X_i$	Sample regression line $\hat{Y}_i = b_0 + b_1 X_i$
• Y_i: actual Y • $\hat{Y}_i$: predicted Y • β_0: Population intercept • β_1: Population slope coefficient • ε_i: Population error term	• Y_i: actual Y • $\hat{Y}_i$: predicted Y • b_0: Sample intercept • b_1: Sample slope coefficient • e_i: Sample error term

Simple Linear Regression(con't)



- Estimate b_0 and b_1 by minimizing the sum of the squared residuals $\sum_{i=1}^{n} e_i^2$.
- $b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = r_{XY}\frac{s_Y}{s_X} = r_{XY}\frac{\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}}, \quad b_0 = \bar{Y} - b_1\bar{X}$
- Interpretation of b_0: When X = 0, the predicted y would be b_0
- Interpretation of b_1: For each unit increase in X, the predicted Y is estimated to increase by b_1, keeping other things constant.

Question 1

Q1

Suppose that $Y = 67+27X$ for all values of X and Y. What can you say about the sample correlation coefficient r between X and Y?

a) $r = -1$
b) $r = 0$
c) $r = 1$
d) The magnitude of r is unknown until it is estimated using sample data

Question 1

Q1. Suppose that Y = 67+27X for all values of X and Y. What can you say about the sample correlation coefficient r between X and Y?

a) r = -1

b) r = 0

c) r = 1

d) The magnitude of r is unknown until it is estimated using sample data

Solution:

- Y = 67+27X ➔ There is a linear relationship between X and Y and |r|=1
- The slope is 27(positive) ➔ positive linear relationship

Hence we conclude that r=1.

Question 2

Q2

Suppose that r = 0, where r is the sample correlation coefficient between X and Y. Then

a) X and Y are not related at all
b) X and Y are very closely related
c) Neither of the above is necessarily true

Question 2

Q2. Suppose that r = 0, where r is the sample correlation coefficient between X and Y. Then

a) X and Y are not related at all

b) X and Y are very closely related

c) Neither of the above is necessarily true

√

Solution:

$r = 0$➔ There is no linear relationship between X and Y.

The correlation coefficient can indicate whether there is a **linear relationship** between independent variable (X) and dependent variable (Y).

"r=0" only implies X and Y are not linearly related, but does not imply X and Y are not related.

Question 3

Q3

You want to predict expenditure on hamburger purchases Y using one of the income measures X or Z. If $r_{XY} = 0.4$, $r_{ZY} = 0.3$ and $r_{XZ} = 0.6$, then we should use

a) X
b) Z
c) Not use this information

Question 3

Q3. You want to predict expenditure on hamburger purchases Y using one of the income measures X or Z. If $r_{XY} = 0.4$, $r_{ZY} = 0.3$ and $r_{XZ} = 0.6$, then we should use

a) X

b) Z

c) Not use this information

Solution:

The magnitude of correlation coefficient can measure the strength of the linear relationship.

- The strength of the linear relationship between Y and X is $|r_{XY}| = 0.4$.
- The strength of the linear relationship between Y and Z is $|r_{ZY}| = 0.3$.

Since $|r_{XY}| > |r_{ZY}|$, X has stronger linear relationship than Z. We choose X.

Question 4

Q4

You want to predict the consumption sauce W using one of the income measures X or Z. Given the information $r_{WX} = 0.6$, $r_{XZ} = 0.9$ and $r_{ZW} = -0.8$, you should use

a) X
b) Z
c) Not use this information

Question 4

Q4. You want to predict the consumption sauce W using one of the income measures X or Z. Given the information $r_{WX} = 0.6$, $r_{XZ} = 0.9$ and $r_{ZW} = -0.8$, you should use

a) X

b) Z ✓

c) Not use this information

Solution:

The magnitude of correlation coefficient can measure the strength of the linear relationship.

- The strength of the linear relationship between W and X is $|r_{WX}| = 0.6$.
- The strength of the linear relationship between W and Z is $|r_{ZW}| = 0.8$.

Since $|r_{WX}| < |r_{ZW}|$, Z has stronger linear relationship than Z. We choose Z.

Question 5

Q5

Suppose we find that r_{XY} is very close to -1 for a sample. That means

a) X and Y have very weak correlation
b) The X and Y values lie very close to a straight line with slope that equals -1
c) Neither of the above

Question 5

Q5. Suppose we find that r_{XY} is very close to -1 for a sample. That means

a) X and Y have very weak correlation

b) The X and Y values lie very close to a straight line with slope that equals -1

c) Neither of the above √

Solution:

r_{XY} is very close to -1 ➔ There is **strong negative** linear relationship between X and Y

➔ The X and Y values lie very close to a straight line with negative slope.

Question 6

Q6. A fuel-oil distribution company has collected data over a number of years in order to determine the statistical relationship between the daily temperature and the consumption of fuel-oil in single family dwellings. Given the temperature, the company would like to be able to predict the consumption of fuel-oil in order to service their customers better. The company draws sample observations which yield the following information:

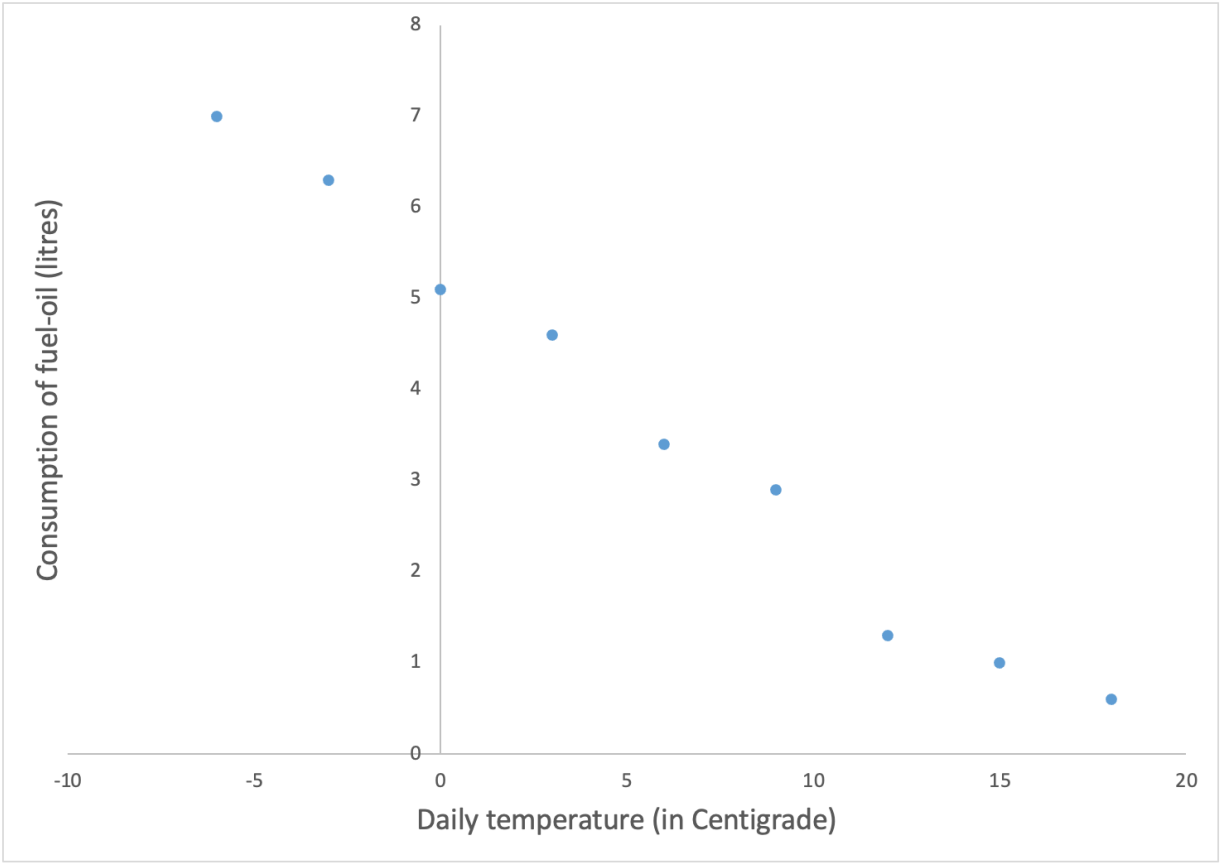
Consumption of fuel-oil (litres)	Daily temperature (in Centigrade)
7.0	-6
6.3	-3
5.1	0
4.6	3
3.4	6
2.9	9
1.3	12
1.0	15
0.6	18

a) Plot the data on a scatter diagram.

Y: Consumption of fuel-oil ; X: Daily temperature

Question 6

a)


Consumption of fuel-oil (litres)
8
7
6
5
4
3
2
1
0
-10
-5
0
5
10
15
20
Daily temperature (in Centigrade)

Question 6

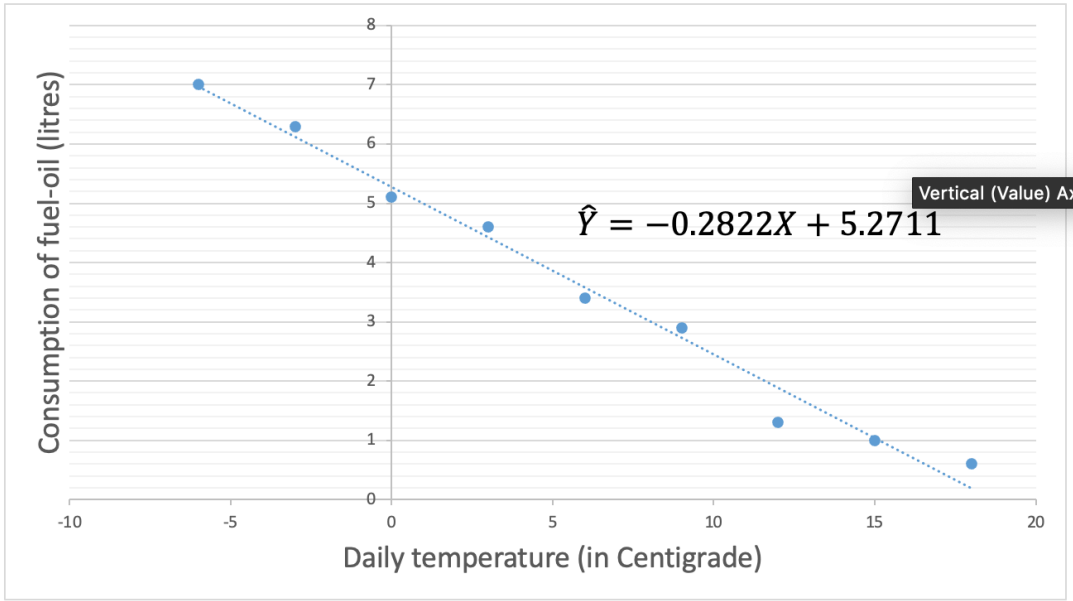
Q6.

b) Estimate the linear regression equation of daily fuel consumption on daily temperature. Superimpose the estimated regression line on the graph constructed in a).

c) Relate the values of the estimated regression coefficients to your diagram.

Solution: b)&c)

$$b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = -0.2822,\ b_0 = \bar{Y} - b_1\bar{X} = 5.2711$$



Question 6

Q6.

d) Predict the level of consumption of fuel when the daily temperature is 7.5c.

e) Measure the strength of the linear relationship between the daily temperature and daily fuel consumption in terms of the sample correlation coefficient. Interpret this value of the coefficient which you obtain.

Solution: d) $\hat{Y} = -0.2822X + 5.2711$

Now X=7.5, then the predicted consumption level

$$\hat{Y} = -0.2822X + 5.2711 = -0.2822 \times 7.5 + 5.2711 = 3.1546$$

Solution e) $r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = -0.992$

➔ The linear relationship between daily temperature and daily fuel consumption is negative and very strong.

4. Students can access the TLQ system during the evaluation period in the following ways:
   a. through a link in an invitation email (a sample is available on the TLQ website ("FAQ" > "Student FAQ" page));
   b. through the course site on Canvas (details are available here);
   c. by logging into the TLQ system directly (https://onlinesurvey.cityu.edu.hk/);
   d. by scanning the TLQ QR Code by smart phones or tablets.