Tutorial 8

Scatter diagram
~ exist, positive/negative, linear/non-linear relationship

Correlation (symmetry)
1. Quantitative - quantitative -> Product-moment/Pearson (values) exact/std
~ very strong/strong/moderate/weak/very weak , positive/negative linear relationship between var
2. Quantitative - ordinal -> Spearman's (rankings) approximation
3. Ordinal - ordinal -> Spearman's (rankings)
~ very high/high/moderate/low/very low degree of agreement/disagreement between the rankings

Regression (model)
Y' = a + bX
a: The estimated DV when IV is 0 unit.
b: The average change in DV when the IV is changed by 1 unit.

* For product-moment correlation coefficient or regression, give the totals will do (no need to show the whole table).

Estimation:
1. Interpolation - within the range of IV -> accurate and reliable
2. Extrapolation - outside the range of IV -> less accurate and not reliable

Coefficient of determination = r^2
About (r^2)x100% of the total variation in Y that is explained by the linear relationship between X and Y. (Correlation)

About (r^2)x100% of the total variation in DV that is explained or accounted for by the total variation in IV. Hence the regression line is considered/not considered as a line of good fit. (Regression)

> 50% < 50%

Q1. As an exercise, a company asked its Store Manager and Purchase Manager to independently rank its eight main suppliers (A, B, C, D, E, F, G and H) in order of value to the company taking discounts and product quality. The two managers ranked the suppliers in order to preference as follows:

Ranking	1	2	3	4	5	6	7	8
Store Manager	E	C	G	H	B	D	A	F
Purchase Manager	E	G	B	D	C	A	H	F

Compute and interpret the Spearman's rank correlation coefficient.

Yu Hong

★ If a particular supplier receives a higher ranking from store manager, he/she will be given a higher ranking from purchase manager.

tutorial 8

(Q1) Let X = store manager, Y = purchase manager

Supplies	A	B	C	D	E	F	G	H
r_x	7	5	2	6	1	8	3	4
r_y	6	3	5	4	1	8	2	7
$d = r_x - r_y$	1	2	-3	2	0	0	1	-3
d^2	1	4	9	4	0	0	1	9

$\sum d^2 = 28$

$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(28)}{8(8^2-1)} = 0.6667$

∴ There is a high degree of agreement between the rankings of store manager and purchase manager for the eight main suppliers - ★

Scanned with CamScanner

Q2. Ten candidates for an administrative post were ranked by the two members of the interviewing panel in the following manner:

Candidate	A	B	C	D	E	F	G	H	I	J
Panel I	4	2	7	1	5	6	9	3	10	8
Panel II	3	2	5	1	4	9	6	7	8	10

Calculate Spearman's rank correlation coefficient and discuss whether it represents a measure of agreement between the two panel members.

Date:

2.	r_x	r_y	$d = r_x - r_y$	d^2
	4	3	1	1
	2	2	0	0
	7	5	2	4
	1	1	0	0
	5	4	1	1
	6	9	-3	9
	9	6	3	9
	3	7	-4	16
	10	8	2	4
	8	10	-2	4

$\Sigma d^2 = 48$

$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6(48)}{10(10^2-1)}$

$= 0.7091$

There is a high degree of agreement between the ranking of the two panel members, Those who are ranked higher in by panel I will also be ranked higher by panel II.

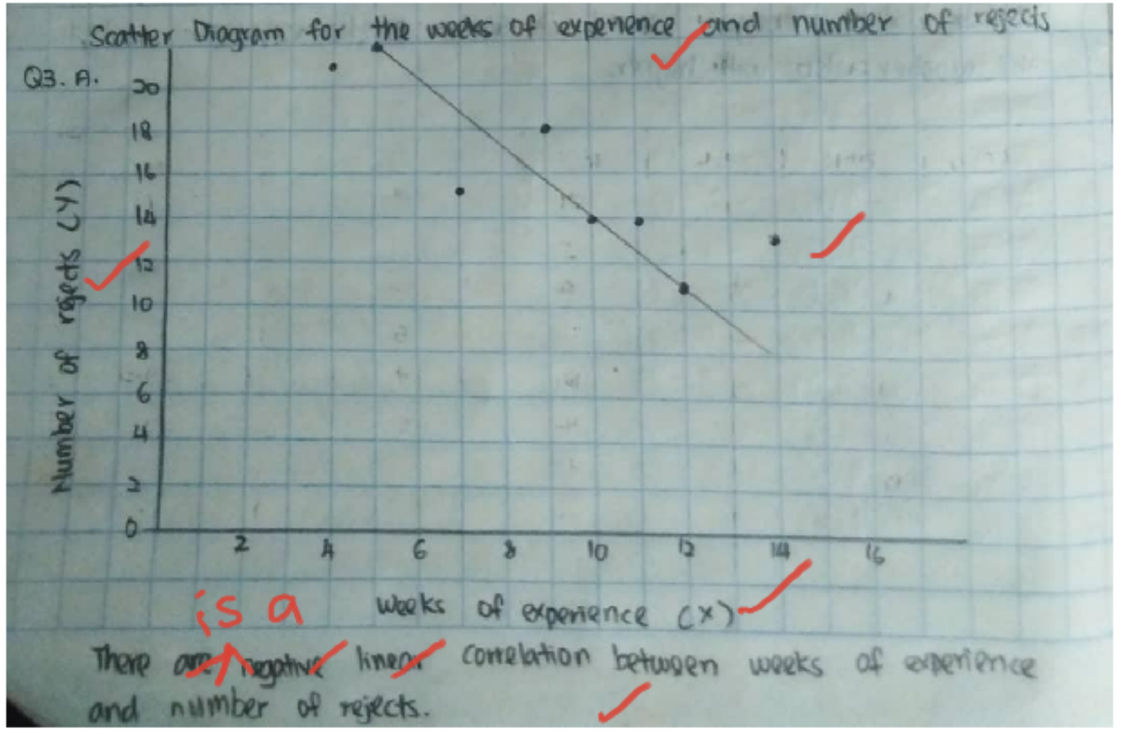
Q3. A sample of eight employees is taken from the production department of a light engineering factory. The data given below relate to the number of weeks experience in the wiring of components, and the number of components which rejected as unsatisfactory last week.

Employee	A	B	C	D	E	F	G	H
Weeks of experience (X)	4	5	7	9	10	11	12	14
Number of rejects (Y)	21	22	15	18	14	14	11	13

(a) Draw and comment the scatter diagram.
(b) Calculate and interpret the product moment correlation coefficient between the two variables.
(c) Compute and interpret the Spearman's rank correlation coefficient between the two variables.

a) & b) Janet ok

c) Aaron ok



There is a negative linear correlation between weeks of experience and number of rejects.

c) ok

X	Y	r_x	r_y	$d = r_x - r_y$	d^2
4	21	1	7	-6	36
5	22	2	8	-6	36
7	15	3	5	-2	4
9	18	4	6	-2	4
10	14	5	3.5	+1.5	2.5 2.25
11	14	6	3.5	2.5	6.25
12	11	7	1	6	36
14	13	8	2	6	36

$\Sigma d^2 = 160.75$ 160.5

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{964.5}{504} = -0.9107$$

$$= 1 - 1.91369$$

$$= -0.9137$$

There is very high degree of disagreement between the rankings of weeks of exp and no. of rejects.

there is a very strong negative linear corelation between weeks of experience and the number of rejects.As the weeks of experience increases, the number of rejects decreases.

Q3.B.

OL.

B.

Weeks of experience, x	Number of rejects, Y	x^2	Y^2	xY
4	21	16	441	84
5	22	25	484	110
7	15	49	225	105
9	18	81	324	162
10	14	100	196	140
11	14	121	196	154
12	11	144	121	132
14	13	196	169	182
$\Sigma X = 72$	$\Sigma Y = 128$	$\Sigma X^2 = 732$	$\Sigma Y^2 = 2156$	$\Sigma XY = 1069$

n=8

$$r = \frac{n\Sigma XY - (\Sigma X \times \Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}} = \frac{8(1069) - (72)(128)}{\sqrt{[8(732) - (72)^2][8(2156) - (128)^2]}}$$

$$= -0.8714$$

There is a very strong negative correlation between weeks of experience and number of rejects. If weeks of experience increases, the number of rejects will decrease.

Q1. A cost accountant has derived the total cost ($000) against output ('000 units) of standard size boxes from a factory over a period of ten weeks, yielding the following data.

Week	1	2	3	4	5	6	7	8	9	10
Output	20	2	4	23	18	14	10	8	13	8
Cost	60	25	26	66	49	48	35	18	40	33

(a) Draw and comment the scatter diagram.
(b) Calculate and interpret the product moment correlation coefficient between the two variables.
(c) Compute and interpret the Spearman's rank correlation coefficient between the two variables.
(d) Comment on the accuracy of the estimates obtained in (b) and (c).

a) & d) Shen Hoi
b) Sean
c) Mavis

4)c) Let X = output, Y = cost

X	20	2	4	23	18	14	10	8	13	8
Y	60	25	26	66	49	48	35	18	40	33
r_X	9	1	2	10	8	7	5	$\frac{3+4}{2}=3.5$	6	$\frac{3+4}{2}=3.5$
r_Y	9	2	3	10	8	7	5	1	6	4
$d=r_X-r_Y$	0	-1	-1	0	0	0	0	2.5	0	-0.5
d^2	0	1	1	0	0	0	0	6.25	0	0.25

$\Sigma d^2 = 8.5$

$r_S = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6(8.5)}{10(10^2-1)} = 0.9485$

There is a very high degree of agreement between the rankings of output and total cost, indicates that a higher output will have a higher total cost.
more

Tutorial 8

Question 4 4B

X= OutPut
Y=Cost
N= Week
n=no of week

$\Sigma X = 120$ $\Sigma X^2 = 1866$ $n = 10$
$\Sigma Y = 400$ $\Sigma Y^2 = 18200$

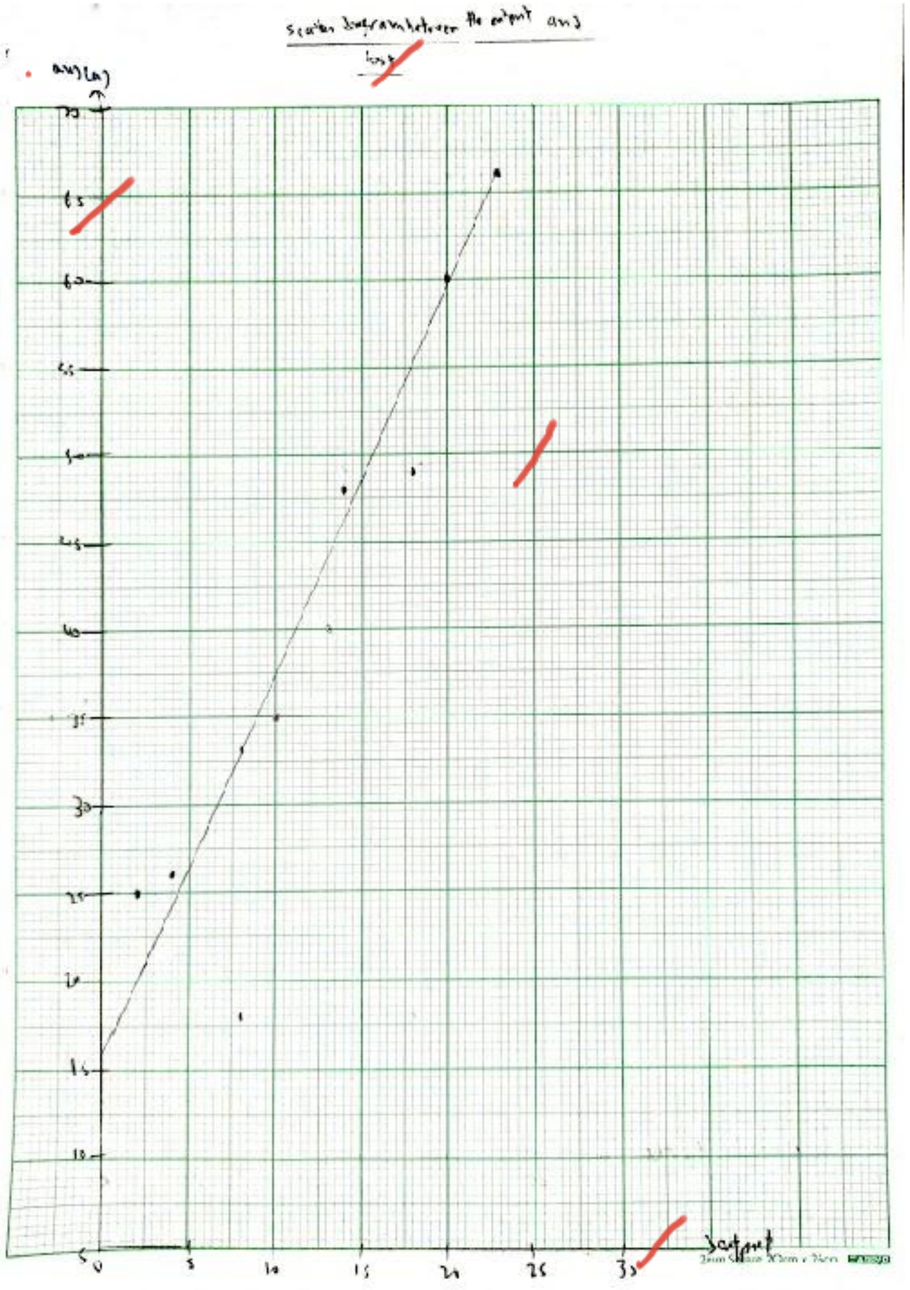
$\Sigma XY = 5704$

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma X)^2}}$$

$$= \frac{10(5704) - (120)(400)}{\sqrt{(10(1866) - (120)^2)(10(18200) - (400)^2)}}$$

$= 0.9338$

There is a very strong positive linear correlation between output and cost. As the output increase, the cost will increase.



9) There is a positive linear correlation between output and cost.

d) r_s is an approximation measurement while r is an exact measurement. r_s can be used to estimate r.

Q5. The following data shows median regional incomes for men aged 21 years and over in full-time employment and average regional house purchase prices for a particular year for the twelve major regions of the United Kingdom.

Region	1	2	3	4	5	6	7	8	9	10	11	12
Median income ($)	57	54	54	51	63	56	52	56	55	55	56	50
House purchase price ($000)	10	9	10	12	15	15	12	11	10	10	11	10

(a) Calculate and interpret the product moment correlation coefficient between the two variables.

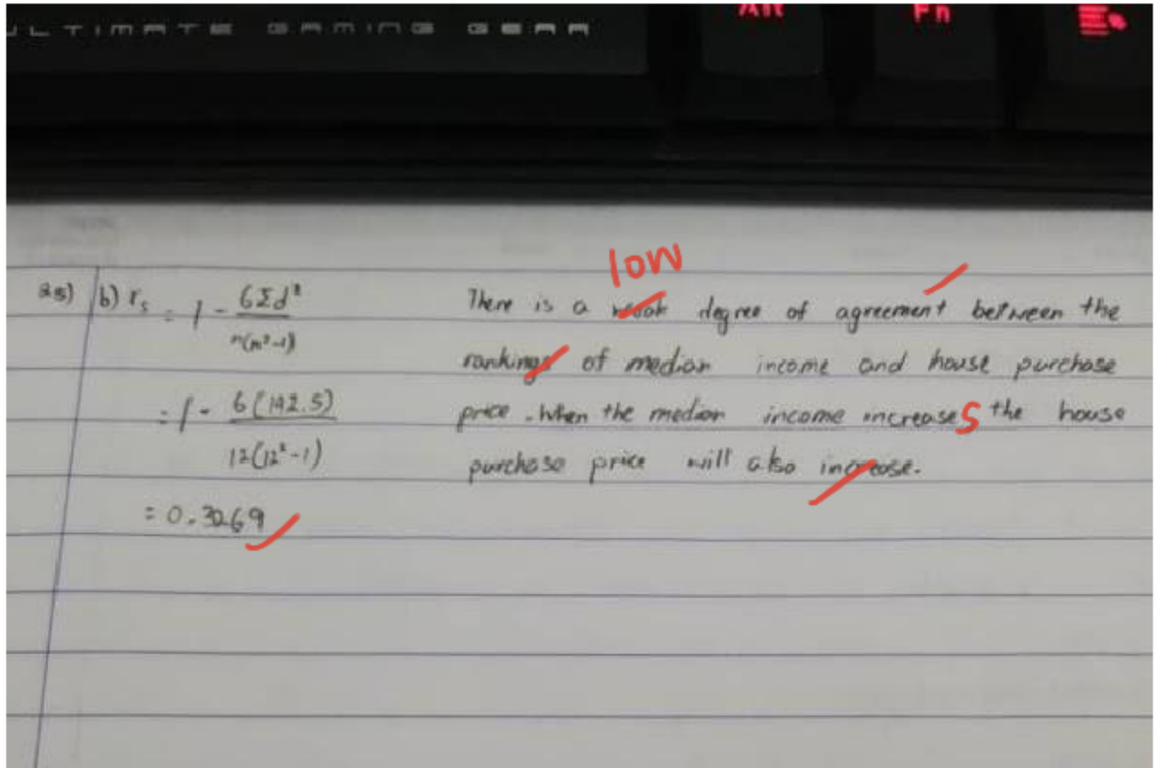
(b) Compute and interpret the Spearman's rank correlation coefficient between the two variables.

a) Yee Hao 8/12

b) Li Yuet 9/12

1 X = $1

1 Y = $1000



3(b) $r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

$= 1 - \frac{6(192.5)}{12(12^2-1)}$

$= 0.3069$

There is a weak degree of agreement between the rankings of median income and house purchase price. When the median income increases the house purchase price will also increase.

low

⑤ Let X = median income ($), Y = House purchase price ($000)

Region	1	2	3	4	5	6	7	8	9	10	11	12
	(11)	(8)	(5)	(3)	(2)	(10)	(3)	(9)	(6)	(8)	(8)	(1)
X	57	54	54	51	63	56	52	56	55	55	56	50
	(6)	(1)	(3)	(9)	(12)	(11)	(10)	(7)	(4)	(5)	(8)	(2)
Y	10	9	10	12	15	15	12	11	10	10	11	10
X^2	3249	2916	2916	2601	3969	3136	2704	3136	3025	3025	3136	2500
Y^2	100	81	100	144	225	225	144	121	100	100	121	100
XY	570	486	540	612	945	840	624	616	550	550	616	500
r_x	11	4.5	4.5	2	12	9	3	9	6.5	6.5	9	1
r_y	4	1	4	9.5	11.5	11.5	9.5	7.5	4	4	7.5	4
$d = r_x - r_y$	7	3.5	0.5	-7.5	0.5	-2.5	-6.5	1.5	2.5	2.5	1.5	-3
d^2	49	12.25	0.25	56.25	0.25	6.25	42.25	2.25	6.25	6.25	2.25	9

$\Sigma X = 659$, $\Sigma Y = 135$, $\Sigma X^2 = 36\,313$, $\Sigma Y^2 = 1561$, $\Sigma XY = 7449$, $\Sigma d^2 = 192.5$

a) $r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}} = \frac{12(7449) - (659)(135)}{\sqrt{[12(36\,313) - (659)^2][12(1561) - (135)^2]}} = 0.4891$

positive

∴ There is a moderate linear correlation between median income and house purchase price. As the median income increases, the house purchase price will increase.

(c) Determine the least squares regression of the house purchase price on the median income for men.
(d) Interpret the regression coefficien *b*.
(e) Estimate the house purchse price in a region where the median regional incomes for men is $ 50 . Comment on the reliability and accuracy of the estimate.
(f) How well does the regression line fit the data?

c) & d) Cecilia
e) & f) Khai Jun

c)House Purchase Price,Y on Median Income For Men
Dependent variable,Y:House Purchase Price
Independent Variable,X :Median Income For Men

$$b = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2};$$ } totals

$$a = \bar{Y} - b\bar{X} \text{ or } a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n}$$

n=12
ΣXY = 7449
ΣX = 659
ΣY:135
ΣX^2 = 36313
b= (12)(7449)-(659)(135) / (12)(36313)-659^2
=423/1475
=0.2868
a=(135/12)-(0.2868)(659/12)
=11.25-15.7501
=-4.5001
Least-squares regression line : Y^1=-4.5001+0.2868X
d)
b=0.2868
The average change in the estimated house purchase price is $286.80 due to a change of $1000 in median income for men.

$1

e) X = 50
Y^1 = -4.5001 + 0.2868(50)
= 9.8399 ($000)

Since the value of X(50) falls within the range of the data set, the estimate is obtained by interpolation technique and hence the estimate is considered as accurate and reliable.

f) "line" - fit
⇓
Coeff of determination

r^2 = 0.4891^2 = 0.2392
=23.92%

About 23.92% at the total variation in house purchase price that is explained or accounted for by the total variation in median income for men. Hence the regression line is not considered as a line of good fit.

Q6. The following table gives the times in seconds for winners in the women's 100-meter freestyle swimming finals in the Summer Olympic Games from 1972 to 2004.

Year	1972	1976	1980	1984	1988	1992	1996	2000	2004
Time	58.6	55.7	54.8	55.9	54.9	54.6	54.5	53.8	53.8

(a) Find the least squares regression line of time on year.

(b) Predict the winning time for the year 2012. Comment on this prediction.

Ze Xuan

Year	1972	1976	1980	1984	1988	1992	1996	2000	2004
Time, Y	58.6	55.7	54.8	55.9	54.9	54.6	54.5	53.8	53.8
X	0	1	2	3	4	5	6	7	8

Independent variable, X = Year (every 4 yrs)

Dependent variable, Y = Time (seconds)

$\Sigma X = 36$ $\Sigma Y = 496.6$ $\Sigma XY = 1959.6$ $\Sigma X^2 = 204$ $n = 9$

$$b = \frac{n(\Sigma XY) - \Sigma X(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{9(1959.6) - 36(496.6)}{9(204) - (36)^2} = -0.4467$$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n} = \frac{496.6}{9} - (-0.4467)\frac{36}{9} = 56.9646$$

$$Y' = 56.9646 - 0.4467X$$

(b) For year 2012, $X = 10$

$Y' = 56.9646 - 0.4467(10)$

$= 52.4976$ seconds

Since the value of X (10) falls outside the range of data set, the estimate is obtained by extrapolation estimate and hence the estimate is considered as less accurate and unreliable.

$b = -0.4467$

→ average change in Y' per every 4 yrs.

Yr	1972	1976	1980	2012
X'	0	4	8	40

$\Rightarrow b' \neq -0.4467$ } $a = 56.9646$, $\hat{Y} = 52.4976$

$= -0.1117$

→ average change in Y' per yr

Q7. The following table shows the average annual earnings of employees in UK from years 2004 to 2010.

Year	2004	2005	2006	2007	2008	2009	2010
Average Annual Earnings ($'000)	59	70	77	87	89	122	137

(a) Find the least squares regression line of average annual earnings on year.
(b) Predict the average annual earnings for the years 2011 and 2012.

Pui Mun

a) independent variable, X = year
dependent variable, Y = average annual earnings ($'000)

Year, X	Y ($'000)	X^2	XY
0	59	0	0
1	70	1	70
2	77	4	154
3	87	9	261
4	89	16	356
5	122	25	610
6	137	36	822
$\Sigma X = 21$	$\Sigma Y = 641$	$\Sigma X^2 = 91$	$\Sigma XY = 2273$

$$b = \frac{n(\Sigma XY) - \Sigma X(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{7(2273) - 21(641)}{7(91) - (21)^2}$$

$$= \frac{2450}{196}$$

$$= 12.5$$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n} = \frac{641}{7} - (12.5)\frac{21}{7}$$

$$= 54.0714$$

$$Y' = a + bX$$

$$Y' = 54.0714 + 12.5X$$

b) For year 2011, $X = 7$

$Y' = 54.0714 + 12.5(7)$

$= 141.5714$ ($)

For year 2012, $X = 8$

$Y' = 54.0714 + 12.5(8)$

$= 154.0714$ ($)

Q8. A random sample of eight drivers insured with a company and having similar car insurance policies was selected. The following table lists their driving experiences and monthly car insurance premiums.

Driving Experience (years)	5	2	12	9	15	6	25	16
Monthly Premium (RM)	64	87	50	71	44	56	42	60

(a) Does the insurance premium depend on the driving experience or does the driving experience depend on the insurance premium? State the dependent and independent variables.

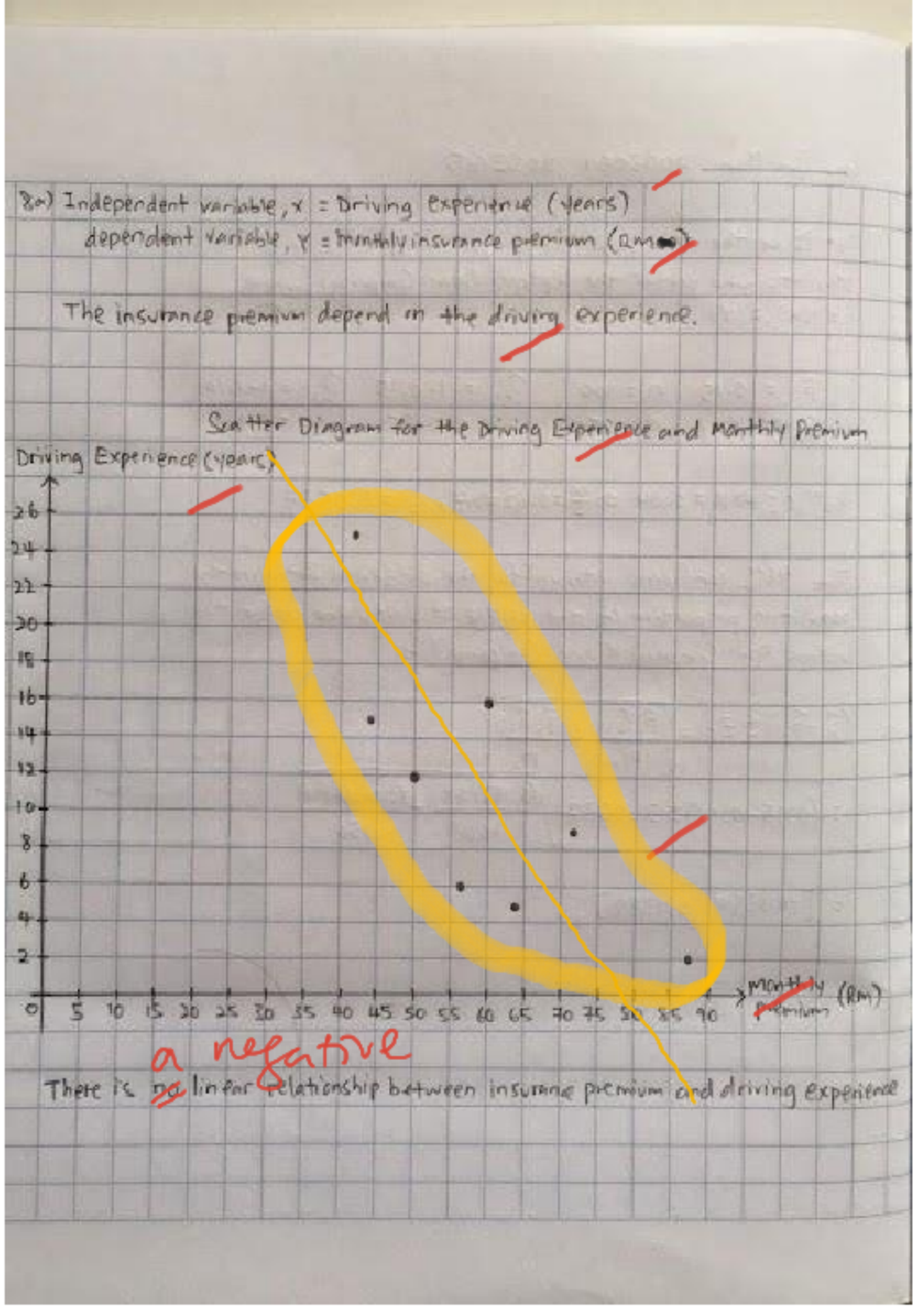
(b) Draw and comment the scatter diagram.

(c) Find the appropriate least squares regression line based on your answer in part (a).

(d) Interpret the meaning of the values of regression coefficients a and b obtained in part (c).

a) & b)Jun Yan OK

c) & d)Kang Hong OK



Q5.

$x = 90$, $y = 474$, $x^2 = 1396$, $xy = 4739$, $n = 8$

The 8(c)

$$b = \frac{n(\Sigma xy) - \Sigma x(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{8(4739) - 90(474)}{8(1396) - (90)^2}$$

$$= \frac{-4748}{3068}$$

$$= -1.5476$$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma x}{n}$$

$$= \frac{474}{8} - (-1.5476)\left(\frac{90}{8}\right)$$

$$= 59.25 + 17.4105$$

$$= 76.6605$$

$$Y' = a + bx$$

$$= 76.6605 - 1.5476x$$

8.(d) For the driver with no experience, ~~that~~ the predicted monthly premium is RM 76.66.

For the driver whose experience increases by 1 year, the monthly premium is expected to decrease by RM 1.55 on average.

(c) Calculate and interpret the correlation coefficient between the two variables.
(f) Calculate and interpret the coefficient of determination.
(g) Predict the monthly car insurance premium for a driver with 10 years of driving experience.
(h) Predict the monthly car insurance premium for a driver with 30 years of driving experience.
(i) Comment on the accuracy of your estimates obtained in parts (g) and (h).

e) & f) Jing Jet OC.
g), h) & i) Jia Jie OC.

e) $r = \frac{n\Sigma xy - \Sigma x(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} = \frac{8(4729) - 90(474)}{\sqrt{[8(1396) - 90^2][8(29\,692) - 474^2]}}$

$= -0.7679$

∴ There is a strong negative correlation between the driving experience and the monthly car insurance premium. As the driving experience of a driver increases, the monthly car insurance premium will decrease.

f) $r^2 = (-0.7679)^2$

$= 0.5897$

$= 58.97\%$

∴ About 58.97% of the total variation in monthly car insurance premium that is explained by the total variation in driving experience. Thus the regression line obtained is considered as fine of good fit.

g) x = 10 years of driving experience
$Y' = 76.6605 - 1.5476(10) = 61.1845$ (CRM)

h) x = 30 years of driving experience
$Y' = 76.6605 - 1.5476(30) = 30.2325$ (CRM)

i) Since the value of $x(10)$ that falls within the range of the data set, the estimate obtained by interpolation technique, hence the estimate obtained $Y' = 61.1845$ is considered accurate or reliable.

Since the value of $x(30)$ that falls outside the range of the data set, the estimate obtained by extrapolation technique, hence the estimate obtained in $Y' = 30.2325$ is considered as not accurate or unreliable.

Q9. The following table gives information on ages and cholesterol levels for a random sample of 10 men.

Age (year)	58	69	43	39	63	52	47	31	74	36
Cholesterol Level	189	235	193	177	154	191	213	165	198	181

(a) State the dependent and independent variables.

(b) Find the least squares regression line of cholesterol level on age.

(c) Interpret the meaning of the values of regression coefficients a and b obtained in part (b).

Chun Wai

Tajuk: Tarikh:

9. (a) Independent variable, X: Age (year)
Dependent variable, Y: Cholesterol Level

(b)

X	58	69	43	39	63	52	47	31	74	36	$\Sigma X = 512$
Y	189	235	193	177	154	191	213	165	198	181	$\Sigma Y = 1896$
XY	10962	16215	8299	6903	9702	9932	10001	5115	14652	6516	$\Sigma XY = 98307$
X^2	3364	4761	1849	1521	3969	2704	2209	961	5476	1296	$\Sigma X^2 = 28110$
											$n = 10$

$$b = \frac{n(\Sigma XY) - \Sigma X(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{10(98307) - 512(1896)}{10(28110) - 512^2}$$

$$= 0.6498$$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n}$$

$$= \frac{1896}{10} - 0.6498\left(\frac{512}{10}\right)$$

$$= 156.3302$$

$Y' = 156.3302 + 0.6498X$

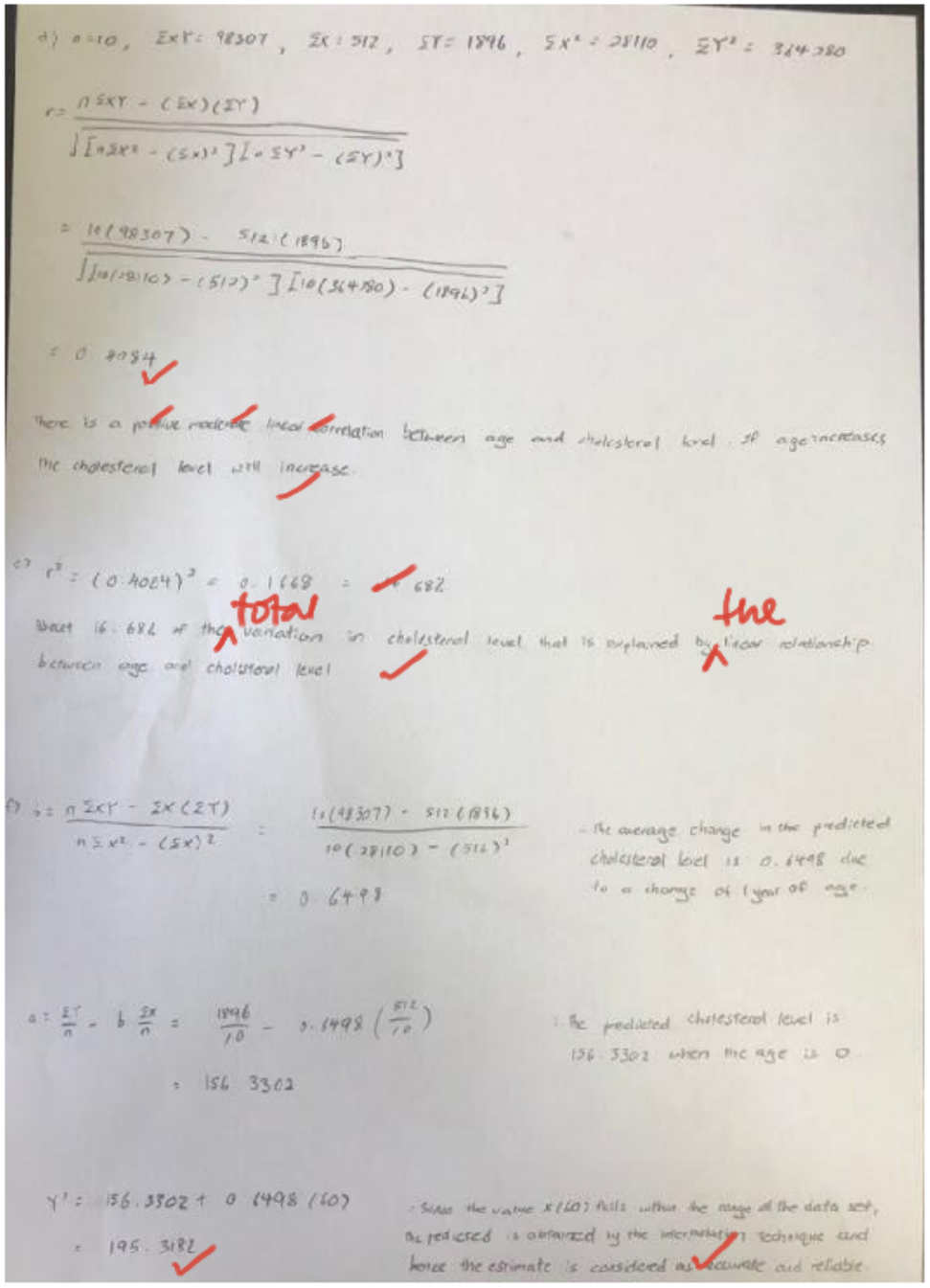
(c) The average change in the estimated cholesterol level is 0.6498 due to a change of 1 year in age.

The estimated cholesterol level is 156.3302 when the age is 0 years.

Jing Xian

(d) Calculate and interpret the correlation coefficient between the two variables.
(e) Calculate and interpret the coefficient of determination.
(f) Predict the cholesterol level of a 60-year-old man.

or .



d) $n=10$, $\Sigma XY = 98307$, $\Sigma X = 512$, $\Sigma Y = 1896$, $\Sigma X^2 = 28110$, $\Sigma Y^2 = 364380$

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{10(98307) - 512(1896)}{\sqrt{[10(28110) - (512)^2][10(364380) - (1896)^2]}}$$

$= 0.4024$

There is a positive moderate linear correlation between age and cholesterol level. If age increases the cholesterol level will increase.

c) $r^2 = (0.4024)^2 = 0.1668 = 16.68\%$

About 16.68% of the total variation in cholesterol level that is explained by the linear relationship between age and cholesterol level

f) $b = \frac{n\Sigma XY - \Sigma X(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{10(98307) - 512(1896)}{10(28110) - (512)^2}$

$= 0.6498$

- The average change in the predicted cholesterol level is 0.6498 due to a change of 1 year of age.

$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n} = \frac{1896}{10} - 0.1498\left(\frac{512}{10}\right)$

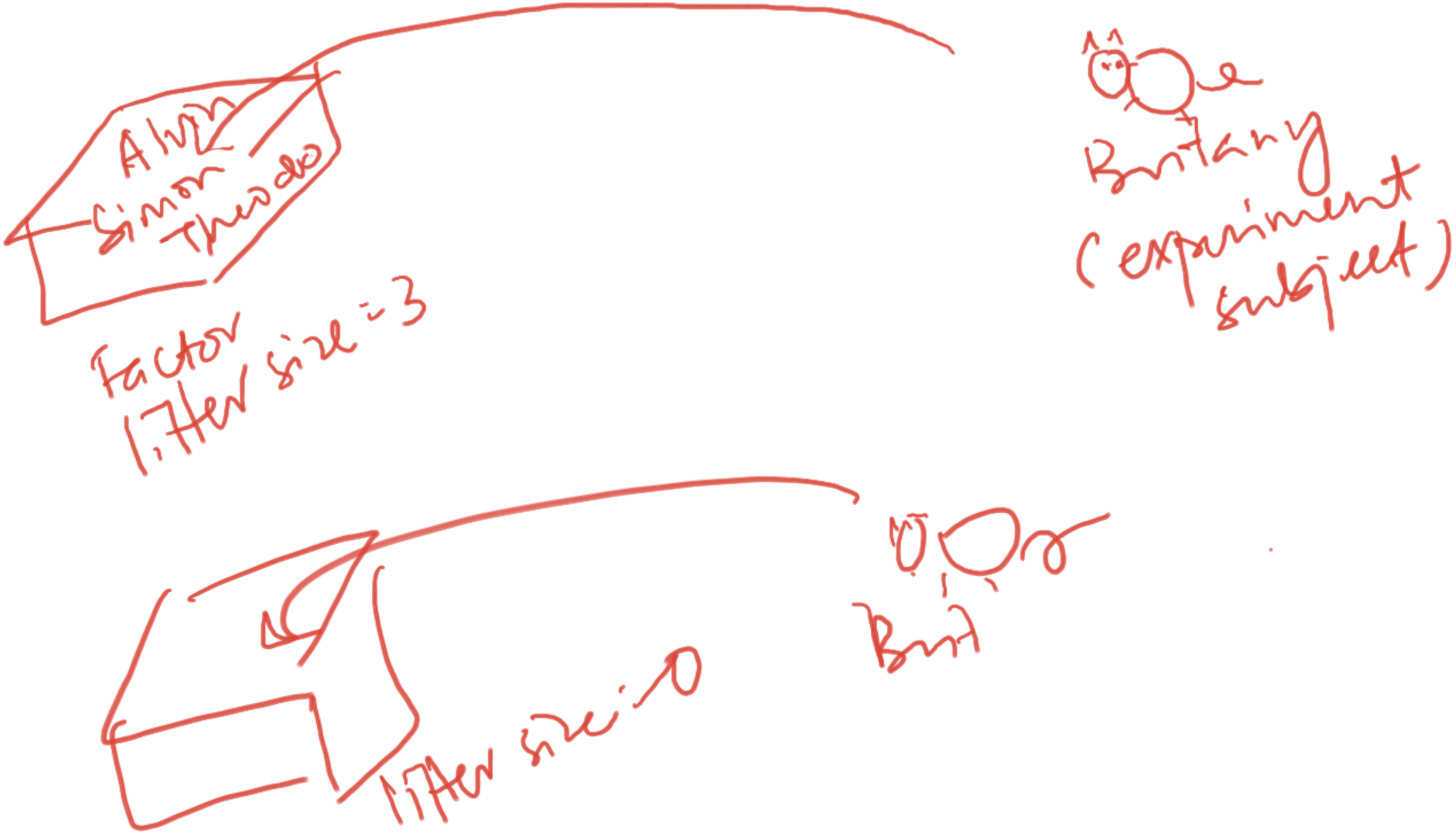
$= 156.3302$

- The predicted cholesterol level is 156.3302 when the age is 0.

$Y' = 156.3302 + 0.6498(60)$

$= 195.3182$

- Since the value $X(60)$ falls within the range of the data set, the predicted is obtained by the interpolation technique and hence the estimate is considered as accurate and reliable.


Alvin
Simon
Theodore
Brittany
(experiment subject)
Factor
Litter size = 3
Litter size = 0
Brit