Correlation & Regression

Ibrahim Rijal - 2106633323

Nomor 1

An insurance company hires an actuary to determine whether the number of hours of safety driving classes can be used to predict the number of driving accidents for each driver.

Nomor 4

Identify the explanatory variable and the response variable

Explanatory variable (Independent): Number of hours of safety driving classes

Response Variable (Dependent): The number of driving accidents for each driver

Nomor 2

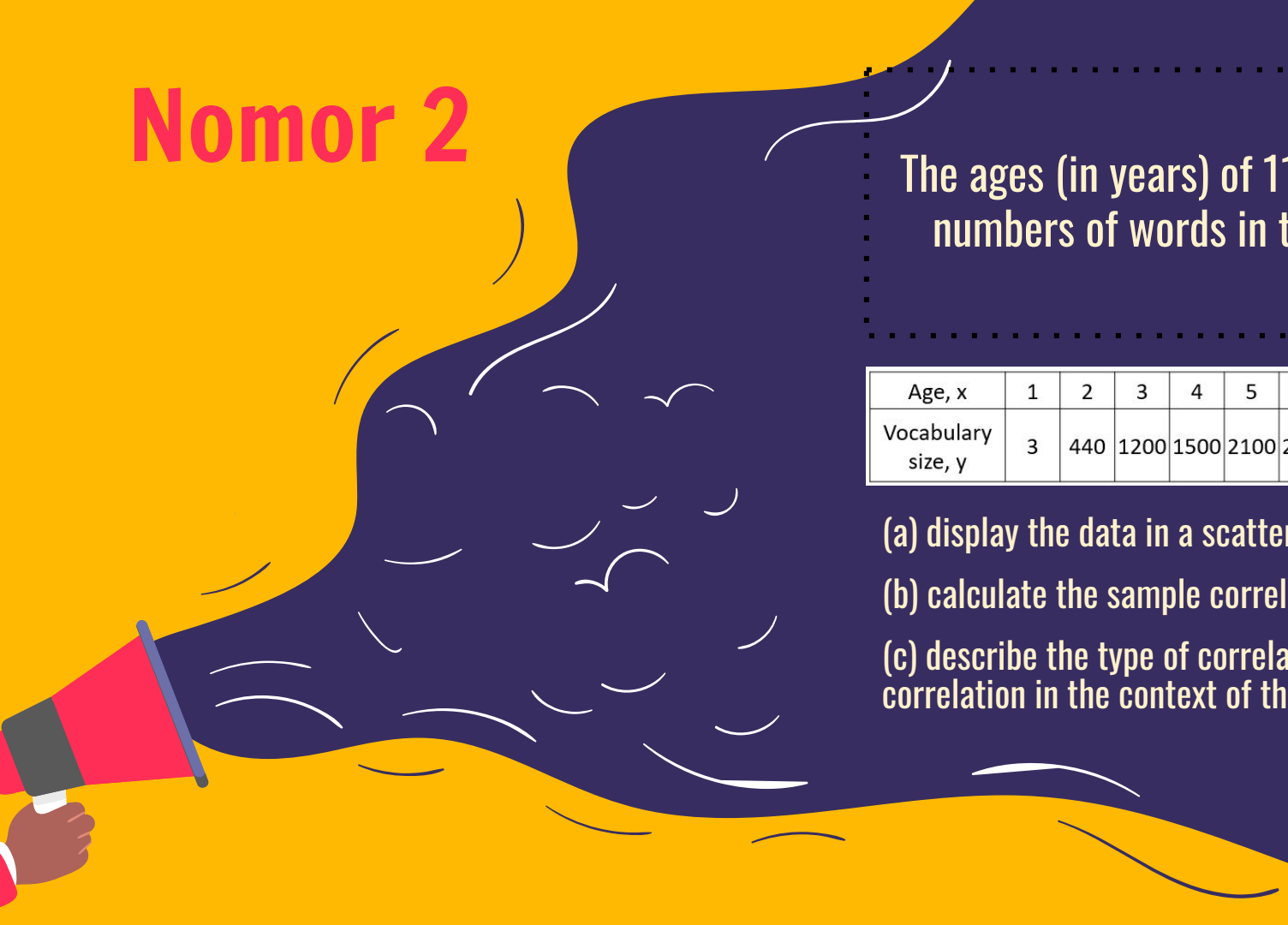
The ages (in years) of 11 children and the numbers of words in their vocabulary

Age, x	1	2	3	4	5	6	3	5	2	4	6
Vocabulary size, y	3	440	1200	1500	2100	2600	1100	2000	500	1525	2500

(a) display the data in a scatter plot

(b) calculate the sample correlation coefficient r

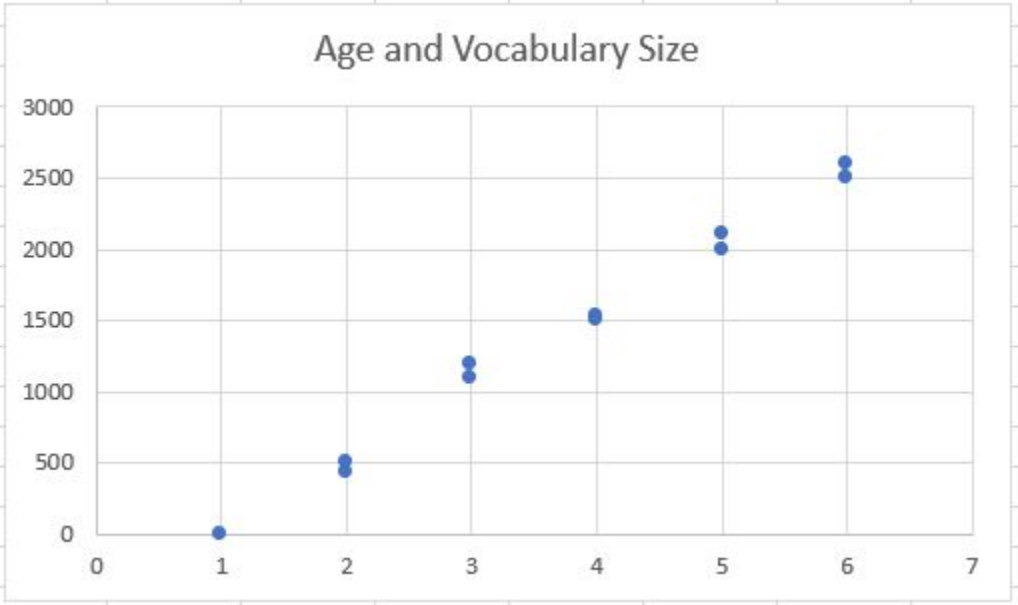
(c) describe the type of correlation and interpret the correlation in the context of the data.



Nomor 2

a) Display the data in a scatter plot

age (x)	vocab size (y)
1	3
2	440
3	1200
4	1500
5	2100
6	2600
3	1100
5	2000
2	500
4	1525
6	2500



Nomor 2

b) Calculate the sample correlation coefficient r

age (x)	vocab size (y)
1	3
2	440
3	1200
4	1500
5	2100
6	2600
3	1100
5	2000
2	500
4	1525
6	2500

=CORREL(B3:B13;C3:C13)

r	0,99643005

Nomor 2

c) Describe the type of correlation and interpret the correlation in the context of the data

From the scatter plot, we can conclude that this is a (very nearly perfect) positive linear correlation, because x increases as y tends to increase.

Interpretation

r = 0.996 shows a strong positive linear correlation, which shows that as children grow older, their vocabulary also expands

Nomor 3

The weights (in pounds) of eight vehicles and the variabilities of their braking distances (in feet) when stopping on a wet surface are shown in the table. At a = 0.05, is there enough evidence to conclude that there is a significant linear correlation between vehicle weight and variability in braking distance on a wet surface?

Weight, x	5890	5340	6500	4800	5940	5600	5100	5850
Variability, y	2.92	2.40	4.09	1.72	2.88	2.53	2.32	2.78

Use Table 11 in Appendix B or perform a hypothesis test using Table 5 in Appendix B to make a conclusion about the correlation coefficient. Compare with using technology

Nomor 3

Use Table 11 in Appendix B or perform a hypothesis test using Table 5 in Appendix B to make a conclusion about the correlation coefficient. Compare with using technology

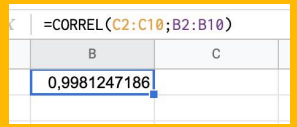
|r| = 0.998 > α

t = r / sqrt((1 - r^2)/(n - 2)) = 38.67

Reject H0

Based on the evidence at a significance level of 50%, it can be concluded that there is a linear relationship between weight and braking distance.

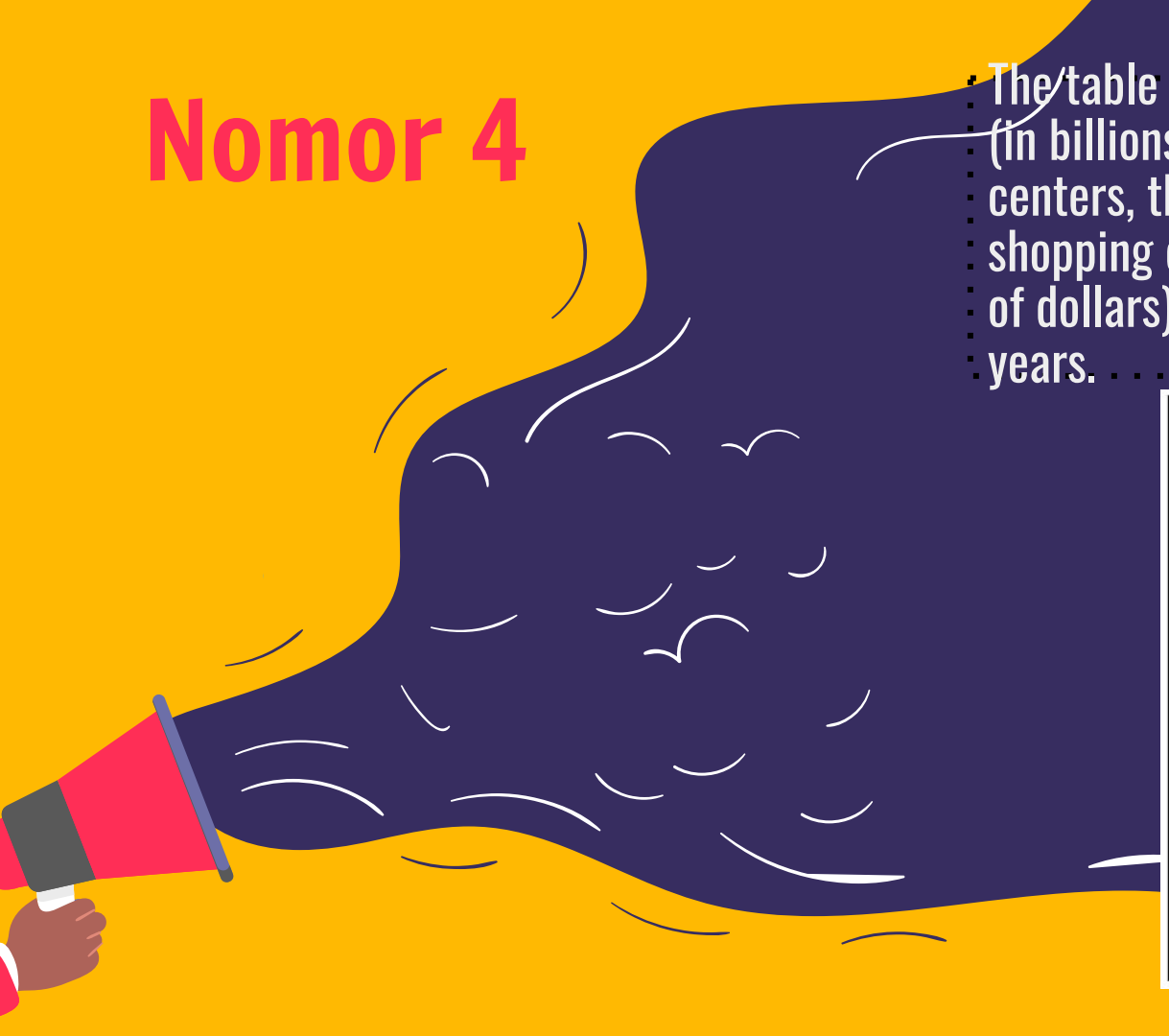
	x	y	x^2	y^2
	4800	1,72	23040000	2,9584
	5100	2,32	26010000	5,3824
	5340	2,4	28515600	5,76
	5600	2,53	31360000	6,4009
	5850	2,78	34222500	7,7284
	5890	2,92	34692100	8,5264
	5940	2,88	35283600	8,2944
	6500	4	42250000	16
Sigma	45020	21,55	255373800	61,0509



Nomor 4

The table shows the total square footages (in billions) of retailing space at shopping centers, the numbers (in thousands) of shopping centers, and the sales (in billions of dollars) for shopping centers for eight years.

Square footage	Sale price
1924	174.9
1592	136.9
2413	275.0
2332	219.9
1552	120.0
1312	99.9
1278	145.0



Nomor 4

a) Find the equation of the regression line for the data.

	x	y	xy	x^2	y^2
	1924	174,9	336507,6	3701776	30590,01
	1592	136,9	217944,8	2534464	18741,61
	2413	275,0	663575	5822569	75625
	2332	219,9	512806,8	5438224	48356,01
	1552	120,0	186240	2408704	14400
	1312	99,9	131068,8	1721344	9980,01
	1278	145,0	185310	1633284	21025
SIGMA	12403	1171,6	2048143	23260365	218717,64

The Equation of a Regression Line

$\hat{y} = mx + b$ where

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \qquad b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

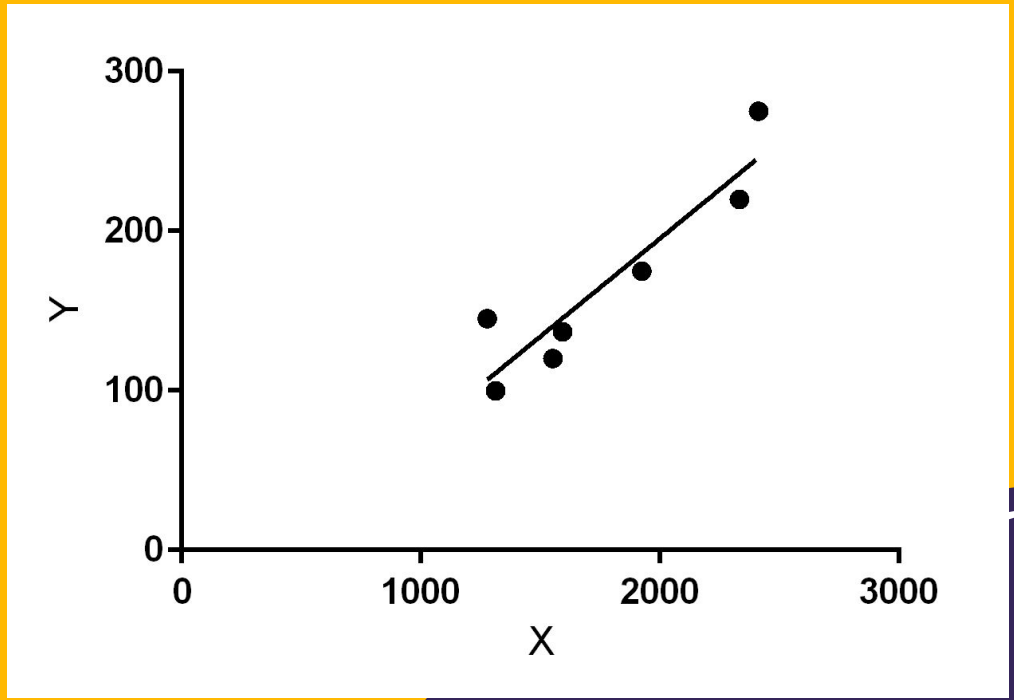
Berdasarkan rumus diatas
m = 0,122696738571002000000
b = -50,02966407

Sehingga y-hat = 0.122x -50.03

Nomor 4

b) Construct a scatter plot of the data and draw the regression line (in 1 figure)

Best-fit values	
Slope	0.1227 ± 0.02265
Y-intercept	-50.03 ± 41.30
X-intercept	407.8
1/Slope	8.150
95% Confidence Intervals	
Slope	0.06445 to 0.1809
Y-intercept	-156.2 to 56.14
X-intercept	-856.3 to 878.2
Goodness of Fit	
R square	0.8544
Sy.x	25.67
Is slope significantly non-zero?	
F	29.33
DFn,DFd	1,5
P Value	0.0029
Deviation from horizontal?	Significant



https://www.graphpad.com/quickcalcs/linear2/

Nomor 4

c) Predict values of y with i. x = 1450 square feet; ii. x = 2175 square feet.

y-hat = 0.122x - 50.03

i. y-hat = 0.122 (1450) - 50.03 = 127,8806069
So, when the square footage is 1450 sq feet, the sale price almost equals to 127,88 thousand dollars

ii. y-hat = 0.122 (2175) - 50.03 = 216,8357423
So, when the square footage is 2175 sq feet, the sale price almost equals to 216,84 thousand dollars

Nomor 5

The table shows the median annual earnings (in dollars) of male and female workers from 10 states in a recent year. The equation of the regression line is ny = 0.939x - 6745.842.

Median annual earnings of male workers, x	41,331	48,389	42,667	43,631	55,116	48,492	37,528	43,425	39,562	40,621
Median annual earnings of female workers, y	30,658	40,019	33,665	31,762	44,937	38,025	28,506	35,691	30,578	32,578

Find:

(a) the coefficient of determination r^2 and interpret the result

(b) the standard error of estimate s_e and interpret the result.

Nomor 5

a) Find the coefficient of determination r^2 and interpret the result

Dari persamaan garis regresi y = 0.939x - 6745.283 didapatkan nilai y, yaitu:
y = 0.939(45.637) - 6745.283 = **36107.3**

Kemudian, dengan menggunakan rumus **margin of error** seperti yang tertera di atas, didapatkan nilai margin of error sebesar **3271**

Maka,
(36107.3 - 3271) < y < (36107.3 + 3271)
32746.3 < y < 39288.3

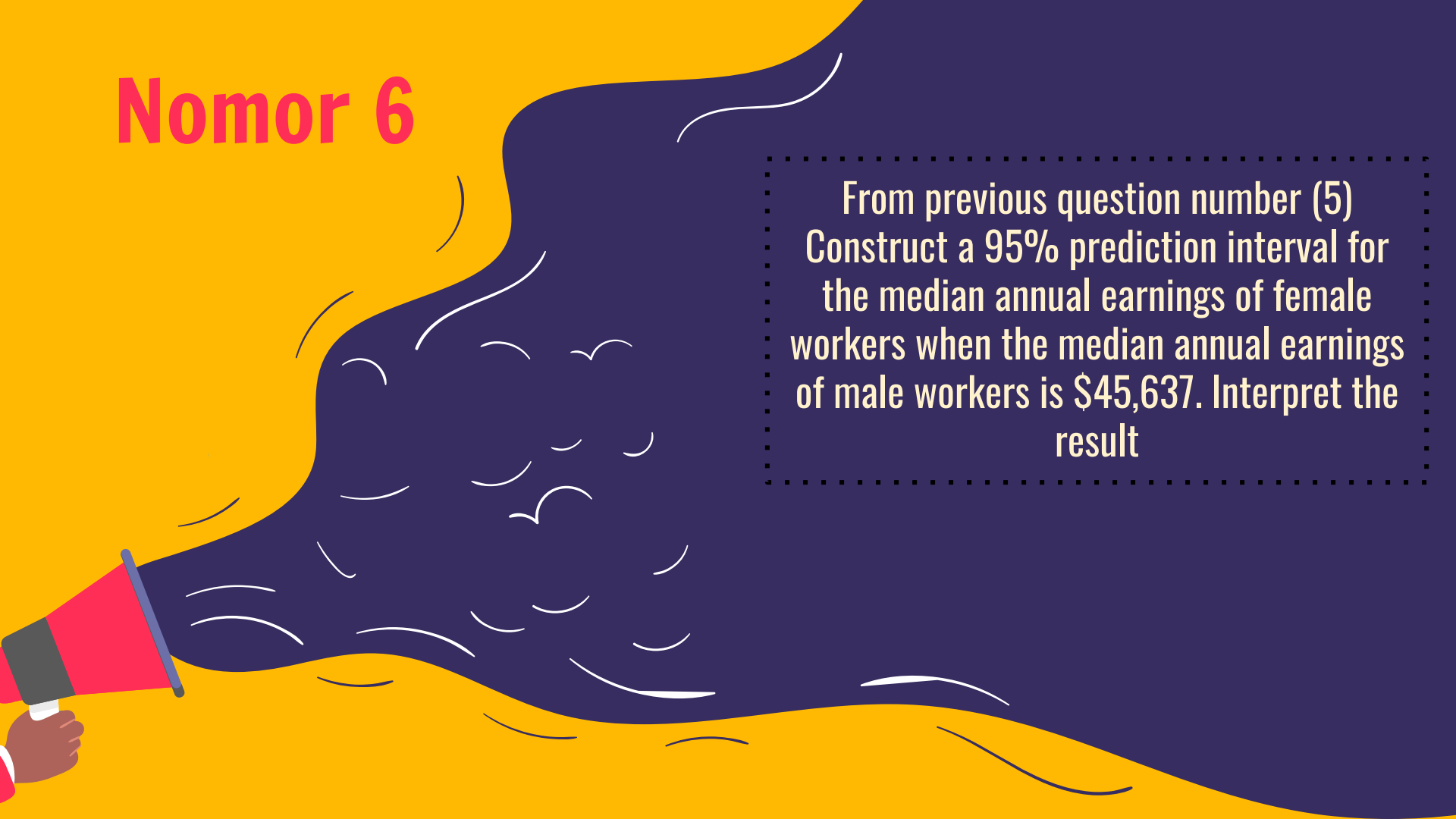
Nomor 5

b) the standard error of estimate se and interpret the result.

$$SEE = \sqrt{\frac{\sum(x_i - \bar{x})}{n-2}}$$

SEE = 1.557

Nomor 6



From previous question number (5) Construct a 95% prediction interval for the median annual earnings of female workers when the median annual earnings of male workers is $45,637. Interpret the result

Nomor 6

$$\text{Margin of error} = Zx\sqrt{\frac{PX\ (1-P)}{N}}$$

Dari persamaan garis regresi y = 0.939x - 6745.283 didapatkan nilai y, yaitu:
y = 0.939(45.637) - 6745.283 = **36107.3**

Kemudian, dengan menggunakan rumus **margin of error** seperti yang tertera di atas, didapatkan nilai margin of error sebesar **3271**

Maka,
(36107.3 - 3271) < y < (36107.3 + 3271)
32746.3 < y < 39288.3

Nomor 7

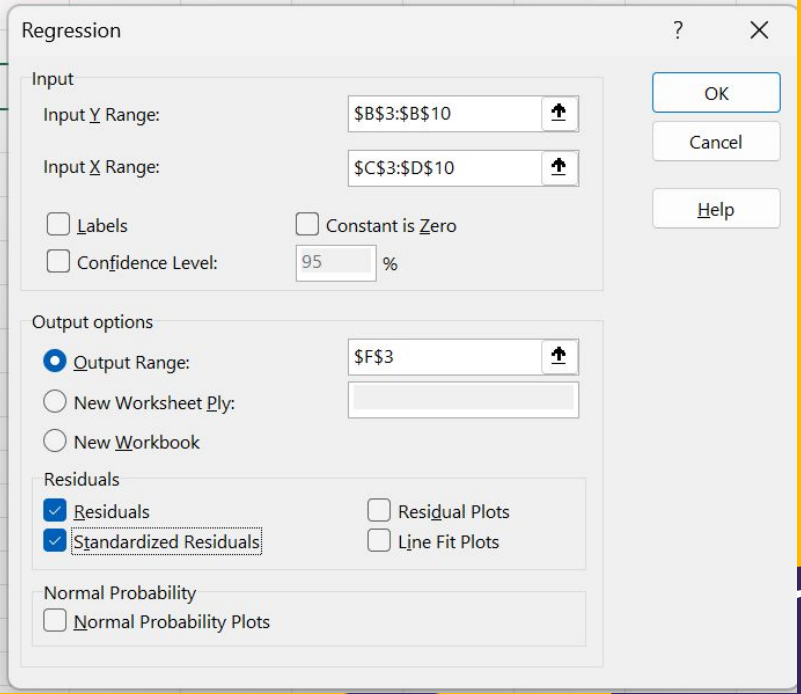
Sales, y	Total square footage, x1	Number of Shopping Center, x2
1032.4	5.7	85.5
1105.3	5.8	87.1
1181.1	6.0	88.9
1221.7	6.1	90.5
1277.2	6.2	91.9
1339.2	6.4	93.7
1432.6	6.5	96.0
1530.4	6.7	98.9

The table shows the total square footages (in billions) of retailing space at shopping centers, the numbers (in thousands) of shopping centers, and the sales (in billions of dollars) for shopping centers for eight years.

Use technology to find (a) the multiple regression equation for the data shown in the table, (b) the standard error of estimate and interpret the result, and (c) the coefficient of determination and interpret the result.

Nomor 7

Sales, y	Total square footage, x1	Number of shopping centre, x2
1032.4	5.7	85.5
1105.3	5.8	87.1
1181.1	6	88.9
1221.7	6.1	90.5
1277.2	6.2	91.9
1339.2	6.4	93.7
1432.6	6.5	96
1530.4	6.7	98.9



Nomor 7

	Coefficients
Intercept	-2075.22702
X Variable 1	20.8958107
X Variable 2	35.0709394

Persamaan regresi:

$$y = 20.90x_1 + 35.07x_2 - 2075.23$$

Regression Statistics	
Multiple R	0.99901274
R Square	0.99802646
Adjusted R Square	0.99723704
Standard Error	8.72131497
Observations	8

$s_e \cong 8.72$

Interpretasi: Standard error of estimate dari hasil penjualan pusat perbelanjaan selama 8 tahun untuk total kuadrat lahan dan jumlah pusat perbelanjaan tertentu adalah $8.72 miliar.

$r^2 \cong 0.998$

Interpretasi: Sekitar 99.8% dari variasi penjualan dapat ditinjau melalui hubungan antara penjualan dengan total kuadrat lahan dan jumlah pusat perbelanjaan. Sekitar 0.2% variasi tidak dapat ditinjau karena beberapa faktor, seperti sampling error, kebetulan, atau variabel tidak diketahui.