HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 7 — Simple linear regression using Excel

Jorge N. Tendeiro

Hiroshima University

Today

We will be using Excel to:

- Compute descriptive statistics.
- Draw plots.
- Perform simple linear regression.
- Make predictions.
- Compute correlations.

From last lecture

We hope that you did the following, as requested at the end of the last lecture:

- Log in to *Moodle*.
- Go to folder "Lecture 7".
- Download two data files: `HOUSE.csv` and `police.csv`.
- Save the two data files to a folder called `Stat` on your Desktop.

Housing data (HOUSE.csv)

Housing data (HOUSE.csv)

> Prices and environment of houses in and around Hiroshima City, which were collected by newspaper advertisements in 1998 and 1999 (number of houses=158; Takahashi, Yanagihara et al., 2000).

Variable	Description
`ID`	ID number
`Price`	Price of single-family house (in $10,000$ yen)
`Period`	Period ($0 = 1998$, $1 = 1999$)
`Area`	Land area (in m^2)
`Size`	Floor area (in m^2)
`JR`	Type of nearest station ($1 =$ JR, $0 =$ otherwise)
`St.time`	Walking time from the nearest station (in minutes)
`Age`	Number of years after being built (in years)
`Distance`	Distance from the center of Hiroshima city (in km)

Housing data (HOUSE.csv)

ID	Price	Period	Area	Size	JR	St.time	Age	Distance
1	3150	0	168	113	0	3	7	10
2	3150	0	177	126	0	4	20	6
3	3500	0	244	77	0	4	21	15
4	6500	0	370	192	0	6	8	13
5	3800	0	148	105	0	7	0	13
6	2890	0	309	128	1	7	8	17

Descriptive statistics, histograms

Descriptive statistics — Sample mean

1. Compute_mean

Jorge Tendeiro

00:31

Link

Procedure

1. Choose an empty cell *to the right of the last column*
   (i.e., **not** under an existing column).

2. Write `=AVERAGE()` and then place the cursor between the empty parentheses.

3. Click on the desired column
   (in our case, `Price`).

4. Press `Enter`.

Conclusion

The mean price of single-family houses is 3291.1 (in 10,000 yen).

For other descriptive statistics

Replace `=AVERAGE()` with a suitable function.

2. Compute_other_statistics

Jorge Tendeiro

01:44

Link

Excel formulas

Statistic	Excel function
variance	`=VAR.P()`
median	`=MEDIAN()`
minimum	`=QUARTILE.INC(, 0)`
1st quartile	`=QUARTILE.INC(, 1)`
2nd quartile (median)	`=QUARTILE.INC(, 2)`
3rd quartile	`=QUARTILE.INC(, 3)`
maximum	`=QUARTILE.INC(, 4)`

Draw a histogram

3. Histogram

Jorge Tendeiro

02:03

Link

Procedure

1. Select the desired column.
2. Go to `Insert` and click on the histogram icon (at 15s).

The histogram is drawn.

You can further edit the histogram if you want:

3. Edit the title (at 26s).
4. Add x and y axis labels (at 42s).
5. Change the bin width, by right-clicking (at 1m26s).

Exercise (1)

1. Calculate the mean, variance, median, and interquartile range of the land area (variable `Area`).

   Hint: Interquartile range = 3rd quartile – 1st quartile
   Use `QUARTILE.INC` and subtraction!!

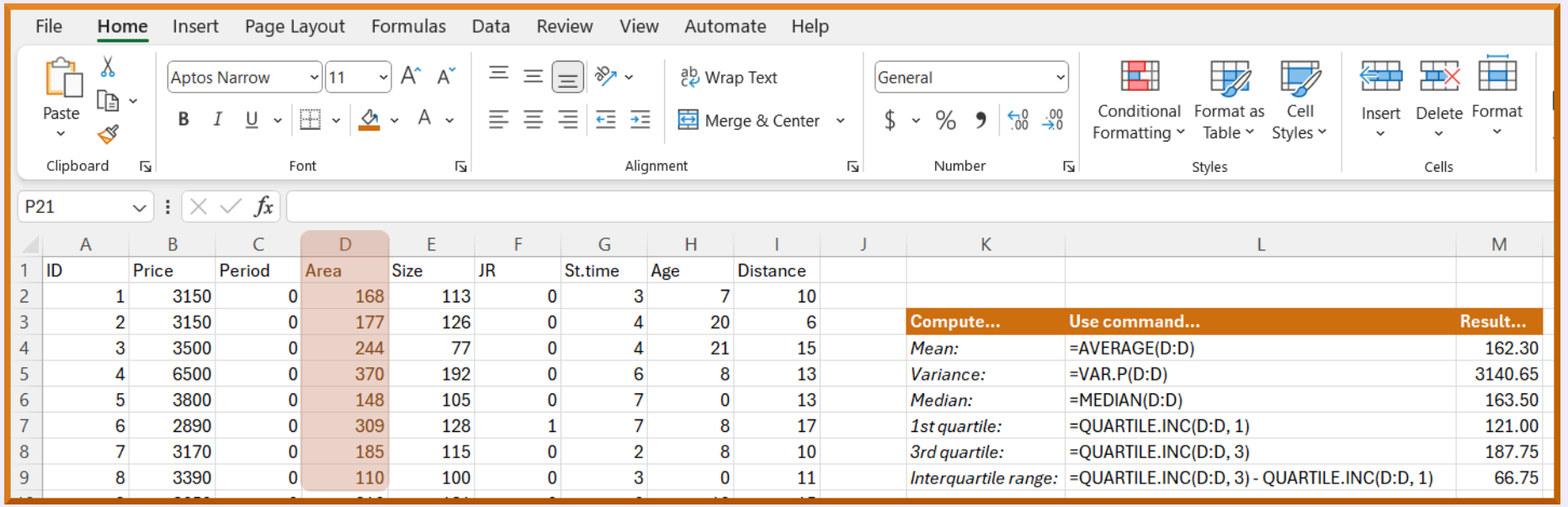
2. Draw a histogram of the land area with a bin width of 30.
   Also, add a title to the graph.

See p. 5 for the data description.

Exercise (1) — ANSWER

1. Calculate the mean, variance, median, and interquartile range of the land area (variable `Area`).

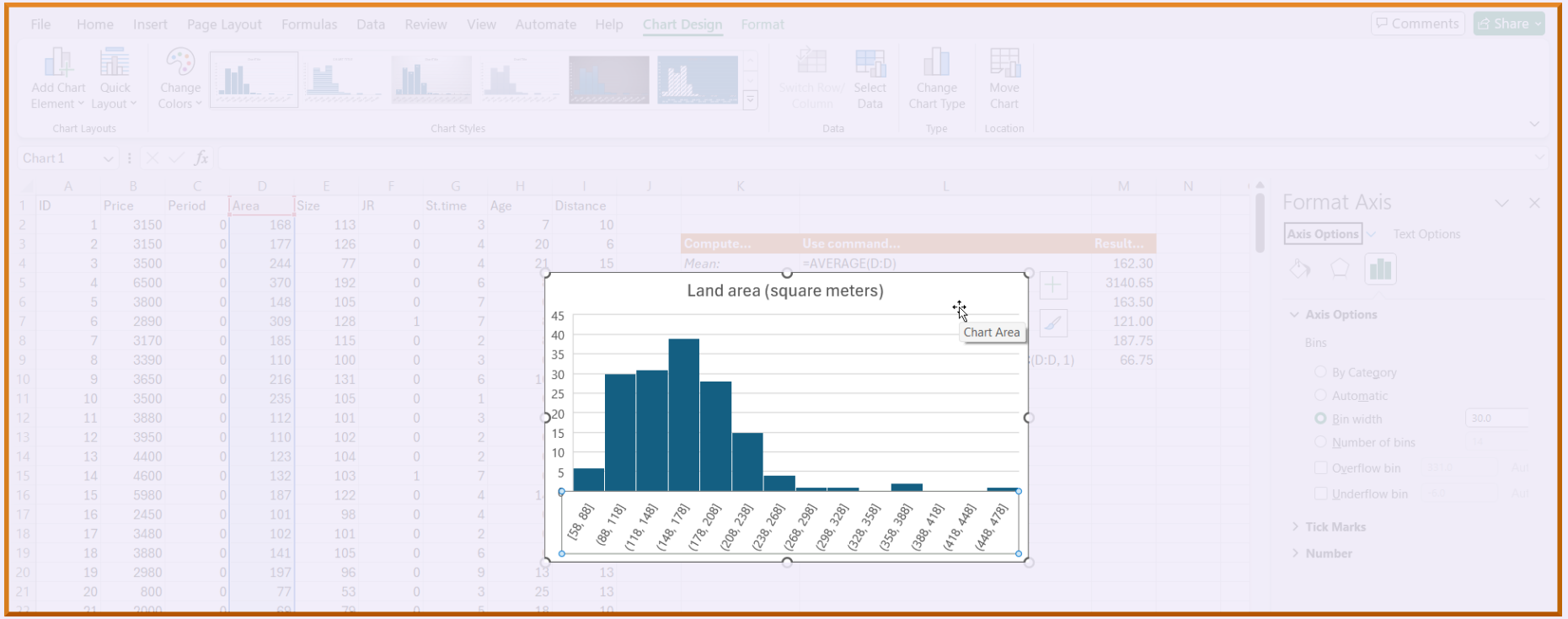
   Hint: Interquartile range = 3rd quartile − 1st quartile
   Use `QUARTILE.INC` and subtraction!!



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	Price	Period	Area	Size	JR	St.time	Age	Distance				
2	1	3150	0	168	113	0	3	7	10				
3	2	3150	0	177	126	0	4	20	6		**Compute...**	**Use command...**	**Result...**
4	3	3500	0	244	77	0	4	21	15		*Mean:*	=AVERAGE(D:D)	162.30
5	4	6500	0	370	192	0	6	8	13		*Variance:*	=VAR.P(D:D)	3140.65
6	5	3800	0	148	105	0	7	0	13		*Median:*	=MEDIAN(D:D)	163.50
7	6	2890	0	309	128	1	7	8	17		*1st quartile:*	=QUARTILE.INC(D:D, 1)	121.00
8	7	3170	0	185	115	0	2	8	10		*3rd quartile:*	=QUARTILE.INC(D:D, 3)	187.75
9	8	3390	0	110	100	0	3	0	11		*Interquartile range:*	=QUARTILE.INC(D:D, 3) - QUARTILE.INC(D:D, 1)	66.75

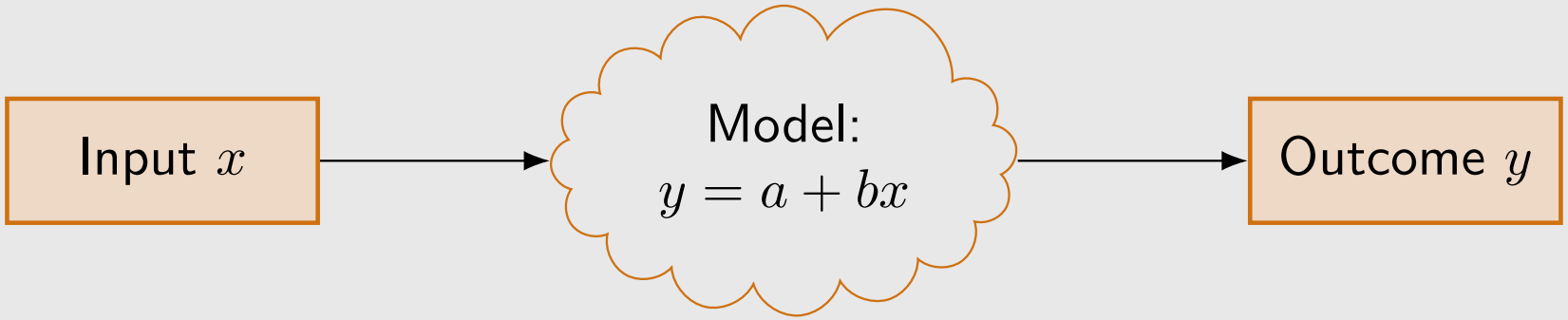
Exercise (1) — ANSWER

2. Draw a histogram of the land area with a bin width of 30.
Also, add a title to the graph.



Simple linear regression using Excel

Simple linear regression (review)



Variables:

- x: Predictor, input, independent variable, explanatory variable.
- y: Response, output, dependent variable, outcome variable.

Goals of simple linear regression:

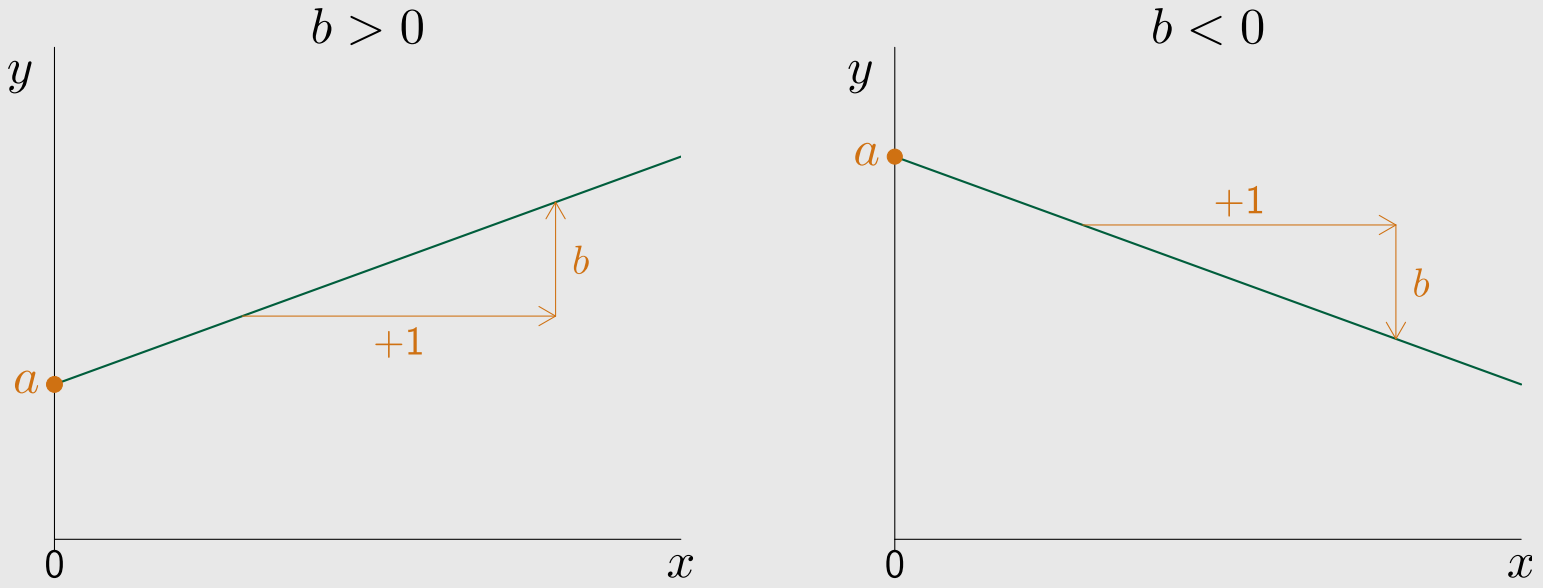
- Quantify the effect of predictor x on outcome y.
- Predict unknown (future) y values for given x values.

Simple linear regression (review)

$$\boxed{y = a + bx}$$

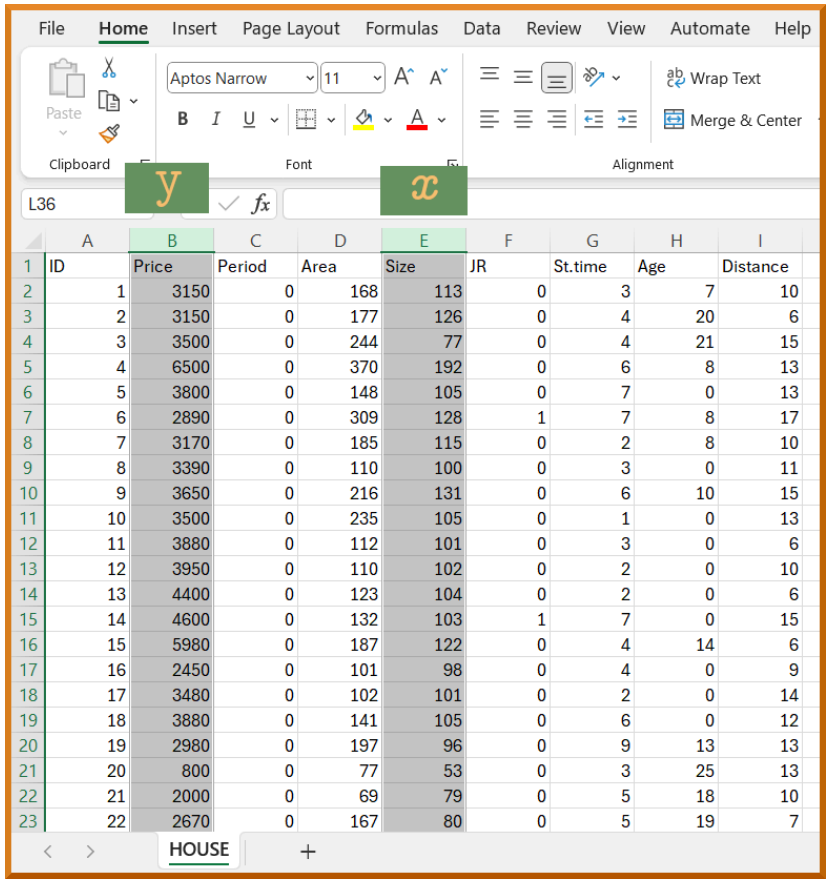
a and b are the model parameters, AKA the regression coefficients.

- a = intercept = predicted value of y when $x = 0$.
- b = slope = amount by which y changes when x increases by 1 unit.



Housing data (HOUSE.csv)

Let's regress `Price` (y variable) on `Size` (x variable).



	A	B	C	D	E	F	G	H	I
1	ID	Price	Period	Area	Size	JR	St.time	Age	Distance
2	1	3150	0	168	113	0	3	7	10
3	2	3150	0	177	126	0	4	20	6
4	3	3500	0	244	77	0	4	21	15
5	4	6500	0	370	192	0	6	8	13
6	5	3800	0	148	105	0	7	0	13
7	6	2890	0	309	128	1	7	8	17
8	7	3170	0	185	115	0	2	8	10
9	8	3390	0	110	100	0	3	0	11
10	9	3650	0	216	131	0	6	10	15
11	10	3500	0	235	105	0	1	0	13
12	11	3880	0	112	101	0	3	0	6
13	12	3950	0	110	102	0	2	0	10
14	13	4400	0	123	104	0	2	0	6
15	14	4600	0	132	103	1	7	0	15
16	15	5980	0	187	122	0	4	14	6
17	16	2450	0	101	98	0	4	0	9
18	17	3480	0	102	101	0	2	0	14
19	18	3880	0	141	105	0	6	0	12
20	19	2980	0	197	96	0	9	13	13
21	20	800	0	77	53	0	3	25	13
22	21	2000	0	69	79	0	5	18	10
23	22	2670	0	167	80	0	5	19	7

Step 1 — Draw a scatterplot of x versus y

4. Scatterplot

Jorge Tendeiro

01:41

Link

Procedure

1. Move column `Size` to the left side of `Price` (at 3s).
   Always place x on the left and y on the right.

2. Draw the scatterplot (at 25s).

3. You can edit the title (at 47s).

4. You can also add labels for the x and y axes (at 56s).

Step 2 — Compute the correlation coefficient r_{xy}

5. Correlation

Jorge Tendeiro

00:52

Link

Procedure

1. Choose an empty cell *to the right of the last column*
   (i.e., **not** under an existing column).
2. Write `=CORREL()` and then place the cursor between the empty parentheses.
3. Click on the x variable, `Size`.
4. Type a comma ",".
5. Click on the y variable, `Price`.
6. Press `Enter`.

Steps 3 & 4 — Draw the simple linear regression line, add equation

6. Add regression line

Jorge Tendeiro

01:39

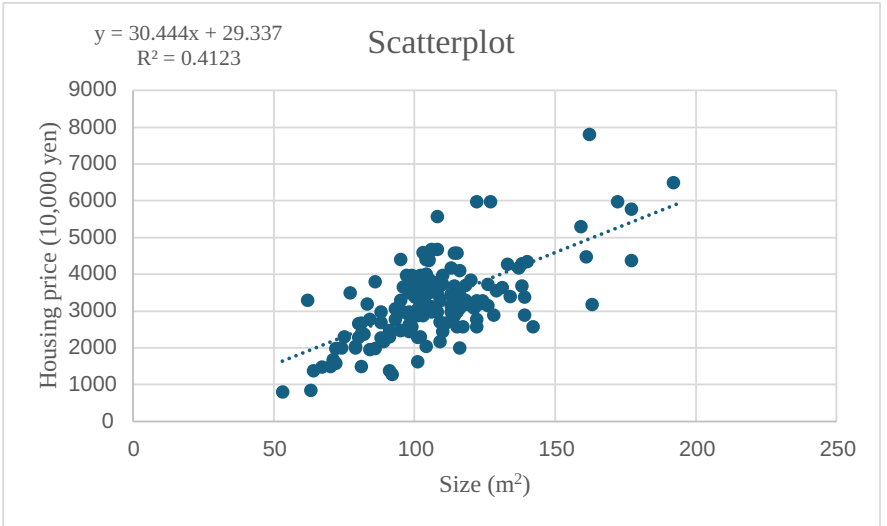
Link

Procedure

1. Select the scatterplot (at 5s).
2. Click on the "Chart Design" menu (at 7s).
3. Click on "Add Chart Element" (at 13s).
4. Select "Trendline > Linear Forecast" (at 21s).
5. Select the regression line, right-click, and choose "Format trendline" (at 32s).
6. Check "Display Equation on chart" and "Display R-squared value on chart" (at 49s).
7. Move the regression equation to make it easier to read (at 1m16s).

Interpret the coefficient of determination

$$\boxed{R^2 = 41.2\%}$$

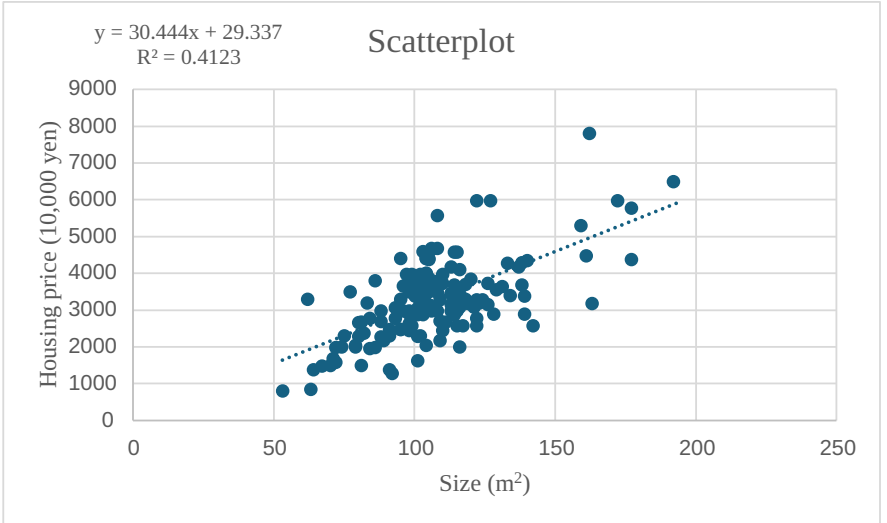


Interpretation

> 41.2% *of the total variance of* `Price` *is explained by predictor* `Size`.

Interpret b, the regression effect of x on y

$$\boxed{\text{Price} = 30.444 \times \text{Size} + 29.337}$$



Interpretation

> *The price of a single-family house is predicted to increase by* $30.444 \times 10,000 = 304,440$ *yen when the floor size increases by* $1m^2$*, on average.*

Step 5 — Make predictions (optional)

7. Prediction

Jorge Tendeiro

01:18

Link

Procedure to predict the price of a house with $150\,m^2$

1. Choose an empty cell *to the right of the last column*
   (i.e., **not** under an existing column).
2. Write `=30.444 * 150 + 29.337`.
3. Press `Enter`.

Interpretation

> *A house of floor area equal to* $150\,m^2$ *is predicted to cost 45,959,370* yen.

Summary

Procedure for running a simple linear regression analysis in Excel:

1. Draw a scatterplot of predictor x versus response y.
2. Compute the correlation coefficient r_{xy}.
3. Draw the simple linear regression line.
   (Ex: the regression line obtained from the least squares method.)
4. Add the linear regression equation and the coefficient of determination.
5. Make predictions (optional).

Exercise (2)

Using the data set in `HOUSE.csv`, set `Age` and `Price` as the explanatory and response variables, respectively, and perform simple linear regression by following the below steps.

1. Draw a scatterplot of `Age` versus `Price`.
   Add a title and axes labels to the plot.
   (Write down the unit of measurement for each axis label.)

2. Calculate the correlation coefficient.

3. Add a regression line to the scatter plot.

4. Add the estimated linear regression equation and the coefficient of determination.

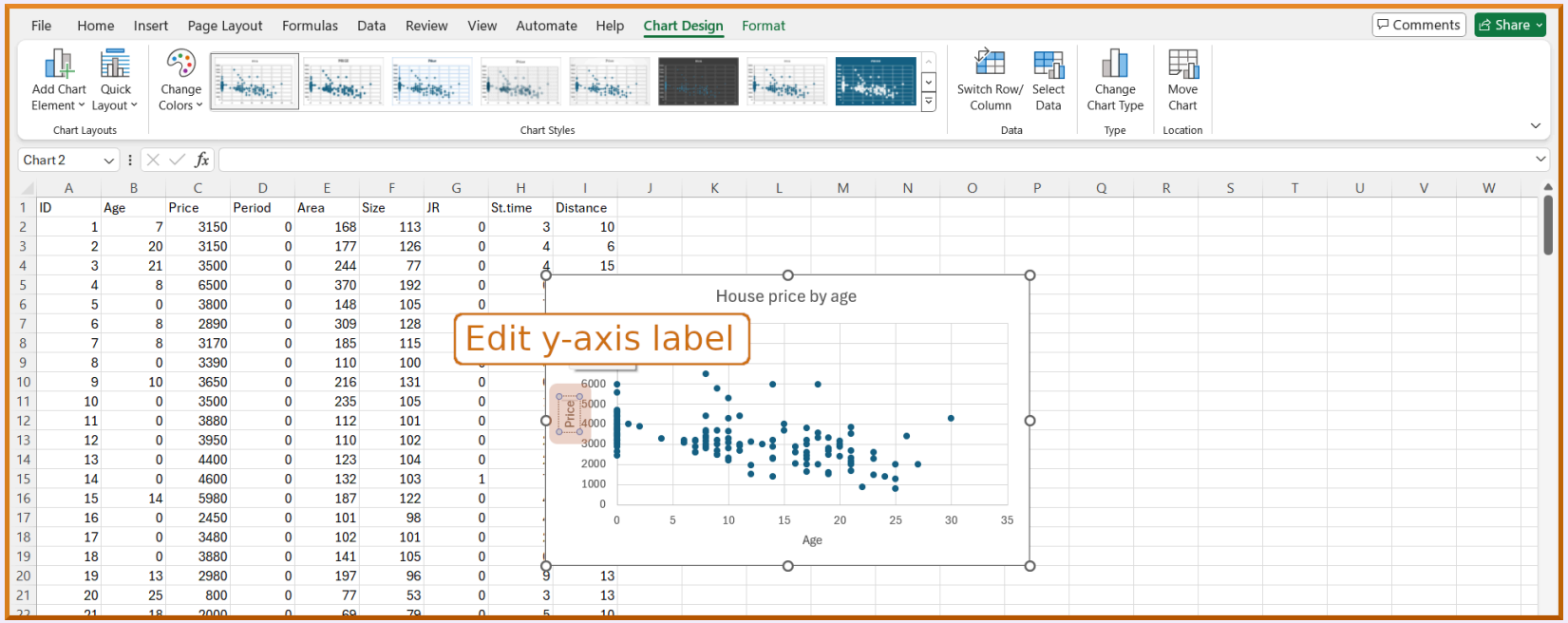
5. Predict `Price` when `Age` is 35 years.

See p. 5 for the data description.

Exercise (2) — ANSWER

Using the data set in `HOUSE.csv`, set `Age` and `Price` as the explanatory and response variables, respectively, and perform simple linear regression by following the below steps.

1. Draw a scatterplot of `Age` versus `Price`.
Add a title and axes labels to the plot.
(Write down the unit of measurement for each axis label.)

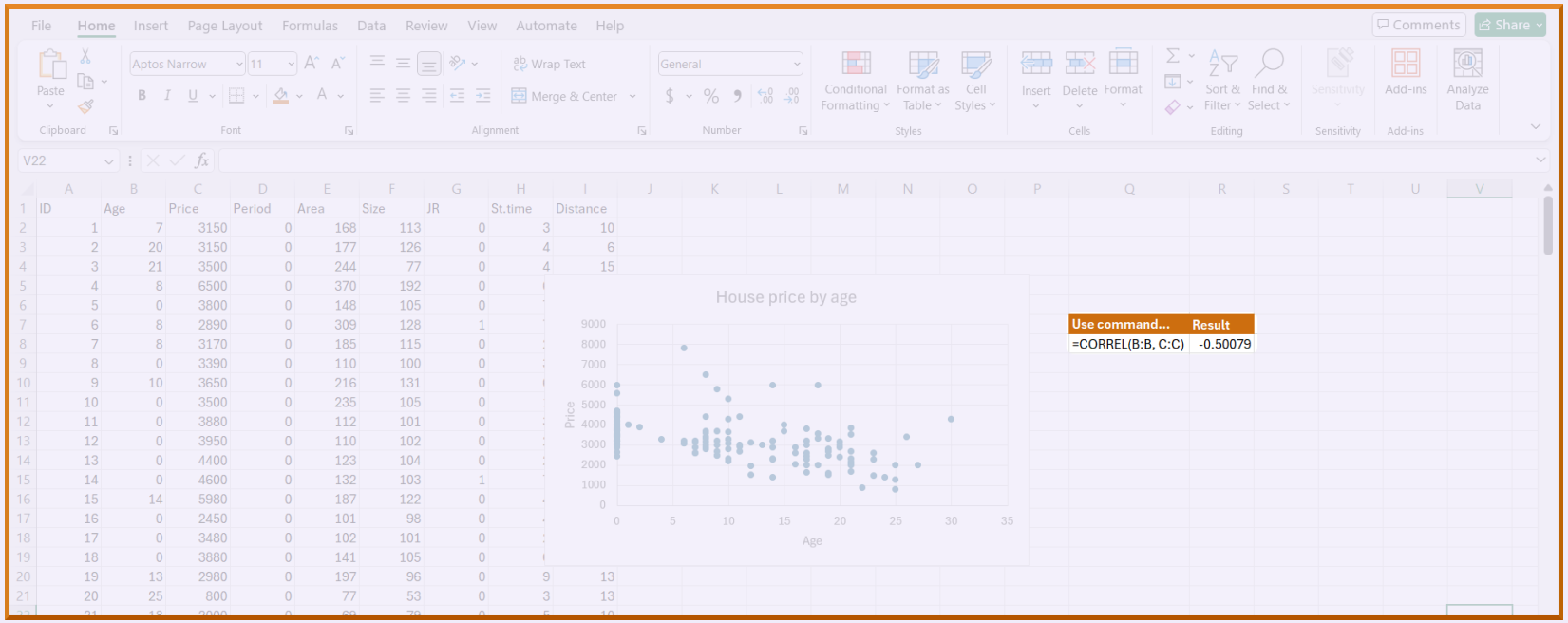


ID	Age	Price	Period	Area	Size	JR	St.time	Distance
1	7	3150	0	168	113	0	3	10
2	20	3150	0	177	126	0	4	6
3	21	3500	0	244	77	0	4	15
4	8	6500	0	370	192	0		
5	0	3800	0	148	105	0		
6	8	2890	0	309	128			
7	8	3170	0	185	115			
8	0	3390	0	110	100			
9	10	3650	0	216	131	0		
10	0	3500	0	235	105	0		
11	0	3880	0	112	101	0		
12	0	3950	0	110	102	0		
13	0	4400	0	123	104	0		
14	0	4600	0	132	103	1		
15	14	5980	0	187	122	0		
16	0	2450	0	101	98	0		
17	0	3480	0	102	101	0		
18	0	3880	0	141	105	0		
19	13	2980	0	197	96	0	9	13
20	25	800	0	77	53	0	3	13

Exercise (2) — ANSWER

Using the data set in `HOUSE.csv`, set `Age` and `Price` as the explanatory and response variables, respectively, and perform simple linear regression by following the below steps.

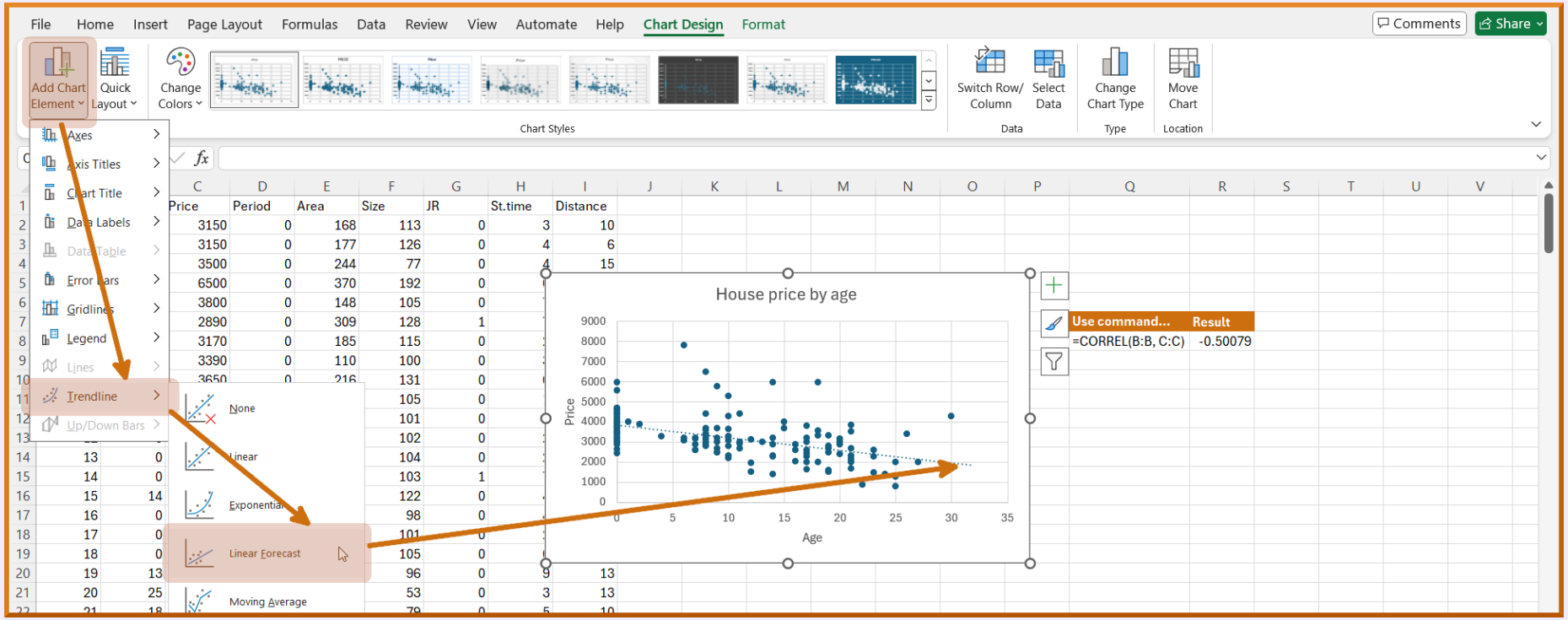
2. Calculate the correlation coefficient.
Add a title and axes labels to the plot.
(Write down the unit of measurement for each axis label.)



Exercise (2) — ANSWER

Using the data set in `HOUSE.csv`, set `Age` and `Price` as the explanatory and response variables, respectively, and perform simple linear regression by following the below steps.

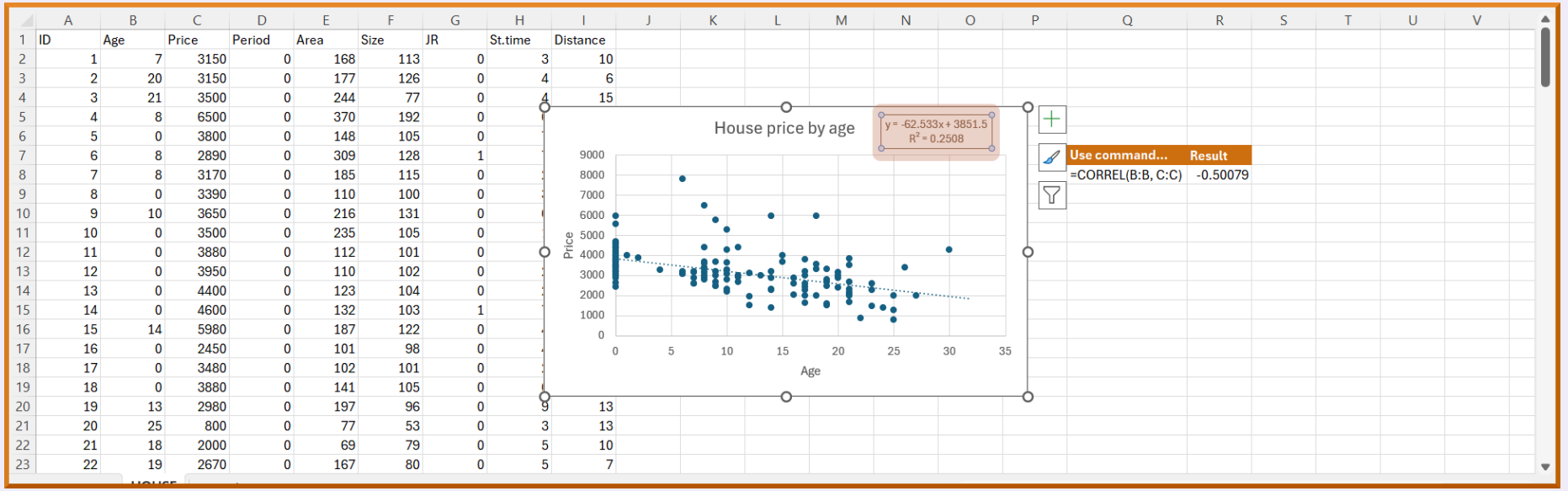
3. Add a regression line to the scatter plot.



Exercise (2) — ANSWER

Using the data set in `HOUSE.csv`, set `Age` and `Price` as the explanatory and response variables, respectively, and perform simple linear regression by following the below steps.

4. Add the estimated linear regression equation and the coefficient of determination.



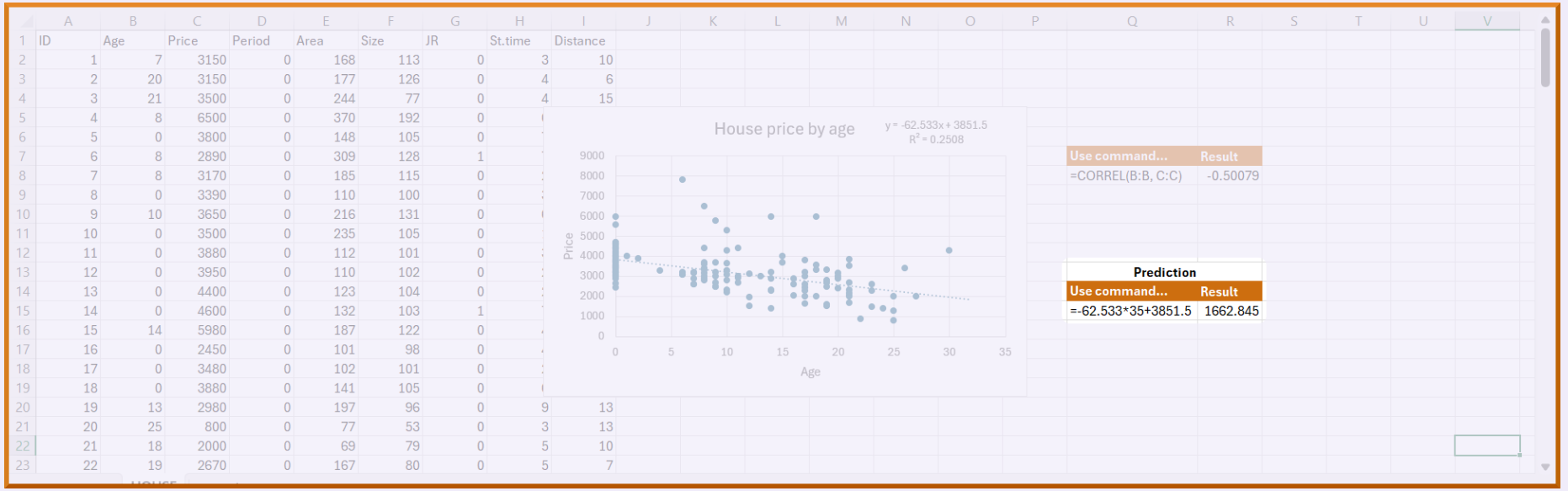
ID	Age	Price	Period	Area	Size	JR	St.time	Distance
1	7	3150	0	168	113	0	3	10
2	20	3150	0	177	126	0	4	6
3	21	3500	0	244	77	0	4	15
4	8	6500	0	370	192	0		
5	0	3800	0	148	105	0		
6	8	2890	0	309	128	1		
7	8	3170	0	185	115	0		
8	0	3390	0	110	100	0		
9	10	3650	0	216	131	0		
10	0	3500	0	235	105	0		
11	0	3880	0	112	101	0		
12	0	3950	0	110	102	0		
13	0	4400	0	123	104	0		
14	0	4600	0	132	103	1		
15	14	5980	0	187	122	0		
16	0	2450	0	101	98	0		
17	0	3480	0	102	101	0		
18	0	3880	0	141	105	0		
19	13	2980	0	197	96	0	9	13
20	25	800	0	77	53	0	3	13
21	18	2000	0	69	79	0	5	10
22	19	2670	0	167	80	0	5	7

Use command...	Result
=CORREL(B:B, C:C)	-0.50079

Exercise (2) — ANSWER

Using the data set in `HOUSE.csv`, set `Age` and `Price` as the explanatory and response variables, respectively, and perform simple linear regression by following the below steps.

5. Predict `Price` when `Age` is 35 years.



ID	Age	Price	Period	Area	Size	JR	St.time	Distance
1	7	3150	0	168	113	0	3	10
2	20	3150	0	177	126	0	4	6
3	21	3500	0	244	77	0	4	15
4	8	6500	0	370	192	0		
5	0	3800	0	148	105	0		
6	8	2890	0	309	128	1		
7	8	3170	0	185	115	0		
8	0	3390	0	110	100	0		
9	10	3650	0	216	131	0		
10	0	3500	0	235	105	0		
11	0	3880	0	112	101	0		
12	0	3950	0	110	102	0		
13	0	4400	0	123	104	0		
14	0	4600	0	132	103	1		
15	14	5980	0	187	122	0		
16	0	2450	0	101	98	0		
17	0	3480	0	102	101	0		
18	0	3880	0	141	105	0		
19	13	2980	0	197	96	0	9	13
20	25	800	0	77	53	0	3	13
21	18	2000	0	69	79	0	5	10
22	19	2670	0	167	80	0	5	7

Exercise (3)

Now consider the `police.csv` dataset.
Perform simple linear regression by setting the number of police officers as the explanatory variable and the number of crimes as the response variable.
Then, predict the number of crimes when the number of police officers is $50,000$.

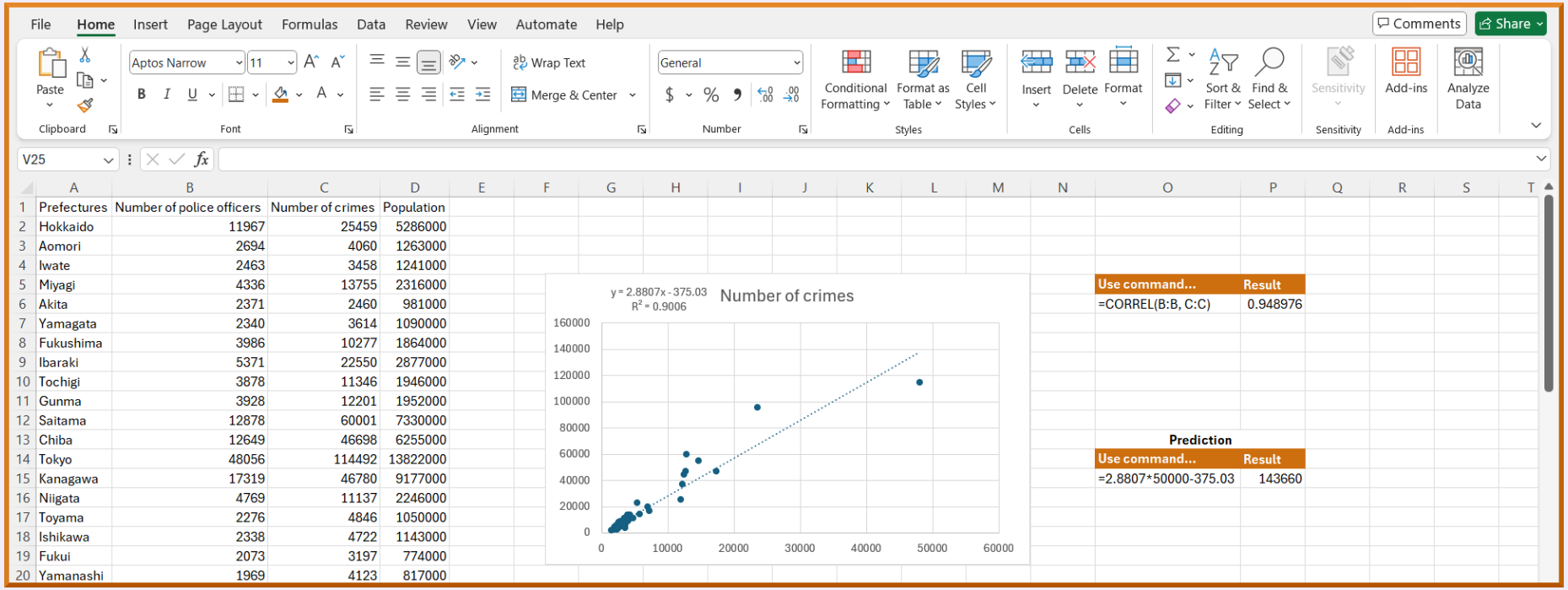
Information on the data file `police.csv`:

Data from each Japanese prefecture with respect to:

- The number of crimes (cases).
  ("White paper", National Police Agency)
- The number of police officers (persons)
  (Ministry of Internal Affairs and Communications)
- Population (persons).
  (Statistics bureau of Japan)

Exercise (3) — ANSWER

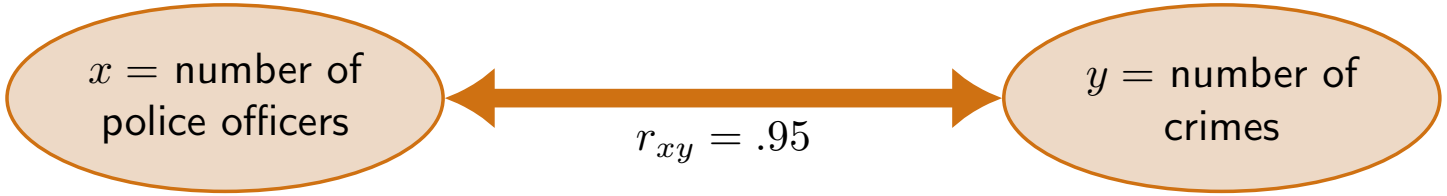
Now consider the `police.csv` dataset.
Perform simple linear regression by setting the number of police officers as the explanatory variable and the number of crimes as the response variable.
Then, predict the number of crimes when the number of police officers is $50,000$.



	A	B	C	D
1	Prefectures	Number of police officers	Number of crimes	Population
2	Hokkaido	11967	25459	5286000
3	Aomori	2694	4060	1263000
4	Iwate	2463	3458	1241000
5	Miyagi	4336	13755	2316000
6	Akita	2371	2460	981000
7	Yamagata	2340	3614	1090000
8	Fukushima	3986	10277	1864000
9	Ibaraki	5371	22550	2877000
10	Tochigi	3878	11346	1946000
11	Gunma	3928	12201	1952000
12	Saitama	12878	60001	7330000
13	Chiba	12649	46698	6255000
14	Tokyo	48056	114492	13822000
15	Kanagawa	17319	46780	9177000
16	Niigata	4769	11137	2246000
17	Toyama	2276	4846	1050000
18	Ishikawa	2338	4722	1143000
19	Fukui	2073	3197	774000
20	Yamanashi	1969	4123	817000

About the correlation coefficient

About the correlation coefficient



> *The correlation is strong, so can we conclude that increasing the number of police officers increases the number of crimes?...*

No!!!

The only thing we can conclude is that variables x and y are linearly associated to each other:

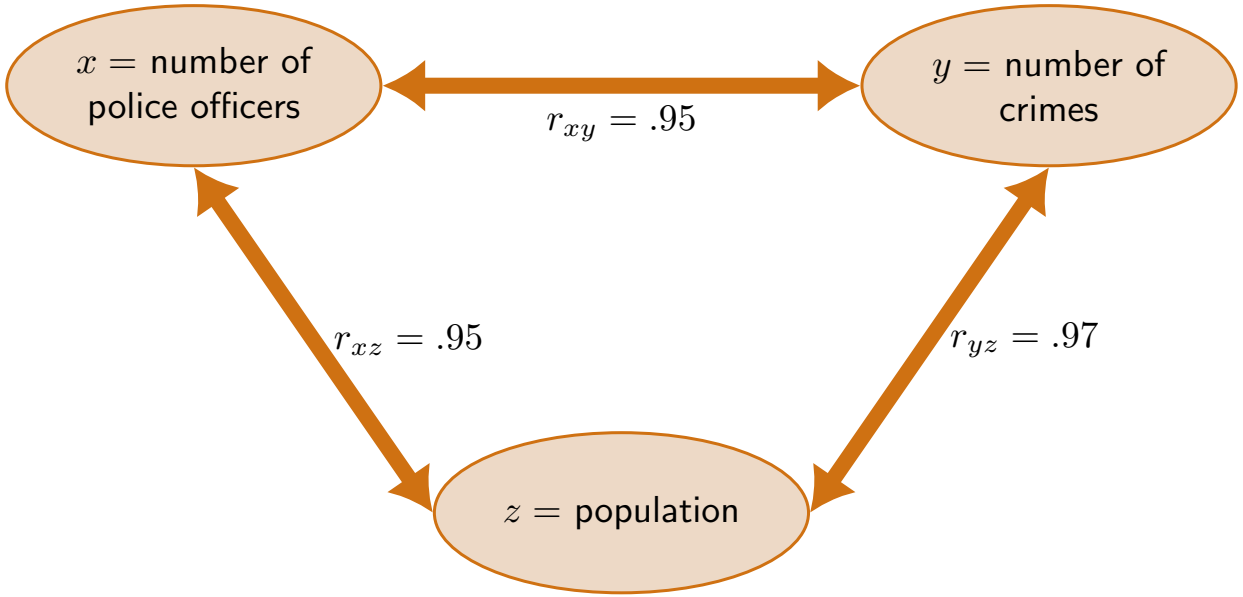
> *y tends to increase when x increases.*

We can **not** conclude that x causes y!!

$$\boxed{\text{correlation} \neq \text{causation}}$$

About the correlation coefficient

Q: Is the correlation between x and y really this large?



A: Maybe not:

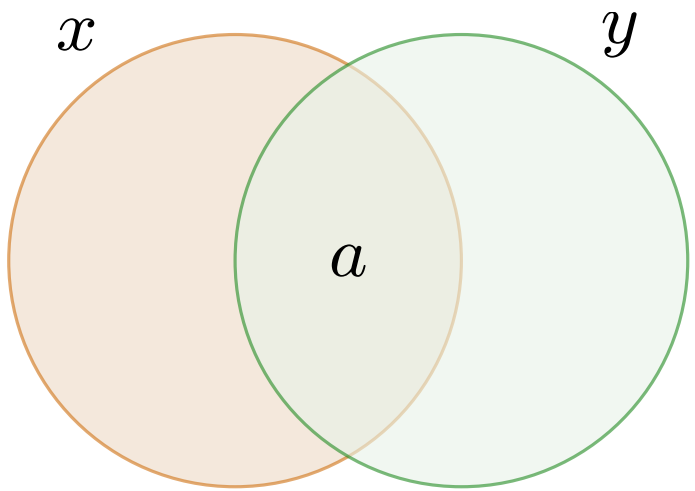
> The relationship between the *number of police officers* and the *number of crimes* is influenced by a 3rd variable, in this case, the *population*.

We say that the correlation between x and y is spurious.

Correlation *versus* partial correlation coefficient

We need to eliminate the effect of the population to interpret the correlation!!
To do that, we can use the partial correlation coefficient.

Let's gain intuition for the partial correlation coefficient.



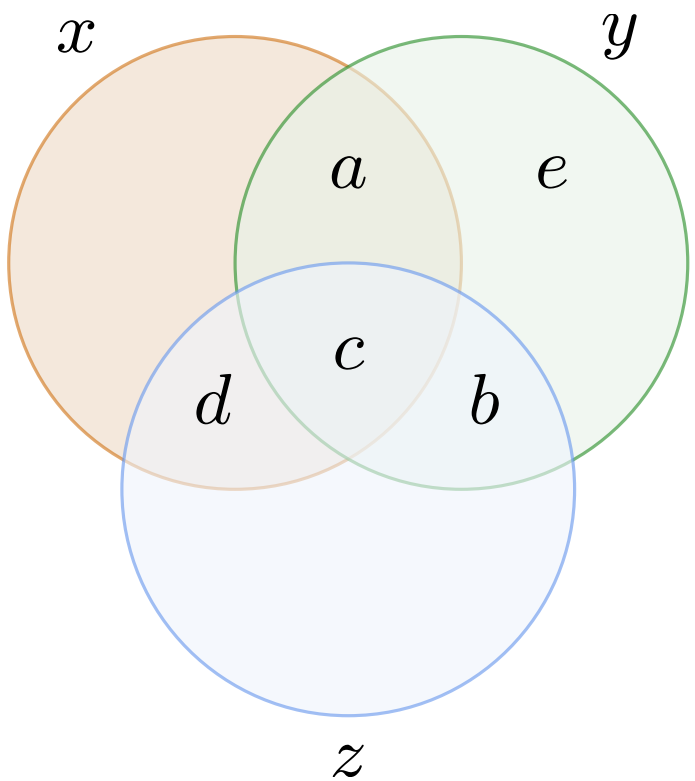
- Area in a circle = variance.
- Assume that all circles have area 1, for simplicity.
- Overlapping area of the two circles = r_{xy}^2.

With this notation,

$$\begin{aligned} r_{xy}^2 &= \text{proportion of } s_y^2 \text{ explained by } x \\ &= \frac{a}{1} \\ &= a. \end{aligned}$$

Correlation *versus* partial correlation coefficient

Now let's consider we have another predictor available, z (like the population in `police.csv`).



Observe that z overlaps with both x and y.

This means that z explains part of the variance of x,

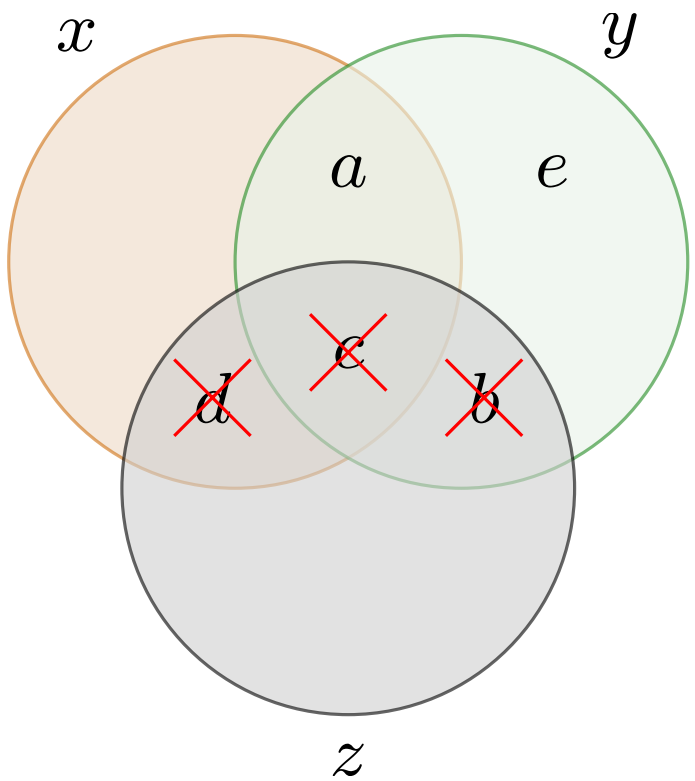
$$r^2_{xz} = c + d$$

and of y,

$$r^2_{zy} = b + c.$$

Therefore, to truly understand the correlation between x and y, we need to control, correct, or adjust for z.

Correlation *versus* partial correlation coefficient



The partial correlation between x and y is given by

$$r_{xy.z} = \frac{a}{a+e}$$

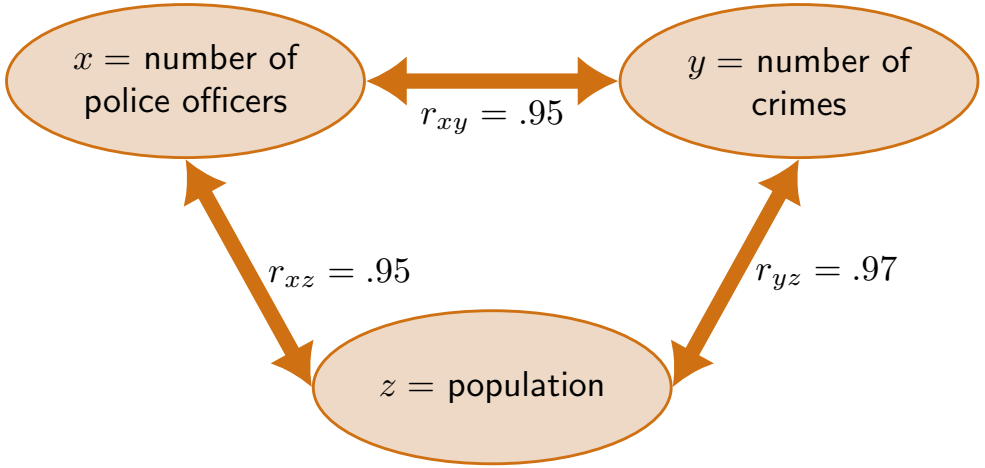
Read:
$r_{xy.z}$ = correlation between x and y, while controlling for z.

Mathematically, the formula to compute $r_{xy.z}$ is

$$\boxed{r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)}}}$$

- r_{xy} = correlation between x and y
- r_{xz} = correlation between x and z
- r_{yz} = correlation between y and z

About the correlation coefficient



$$
\begin{aligned}
r_{xy.z} &= \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)}} \\
&= \frac{.95 - .95 \times .97}{\sqrt{(1 - .95^2)(1 - .97^2)}} \\
&= .38
\end{aligned}
$$

Possible interpretations

1. The correlation between the number of police officers (x) and the number of crimes (y), adjusting for the population (z), is $.38$.

2. The number of police officers (x) *uniquely* explains 38% of the variance of the number of crimes (y) that is not explained by the population (z).

Summary

We learned to use Excel to:

- Compute descriptive statistics.
- Draw plots.
- Perform simple linear regression.
- Make predictions.
- Compute correlations.

We also learned how to adjust correlations for a third variable using the partial correlation coefficient.

To do before the next lecture

Before lecture 8:

- Log in to *Moodle*.
- Go to folder "Lecture 8".
- Download two data files: `seiseki.csv` (download from Lecture 5 folder, if you don't have it) and `iris.csv`.
- Save the two data files to a folder called `Stat` your Desktop.

In the next lecture, we will learn about data analysis methods using RStudio (Cloud).