

CORRELATION & REGRESSION



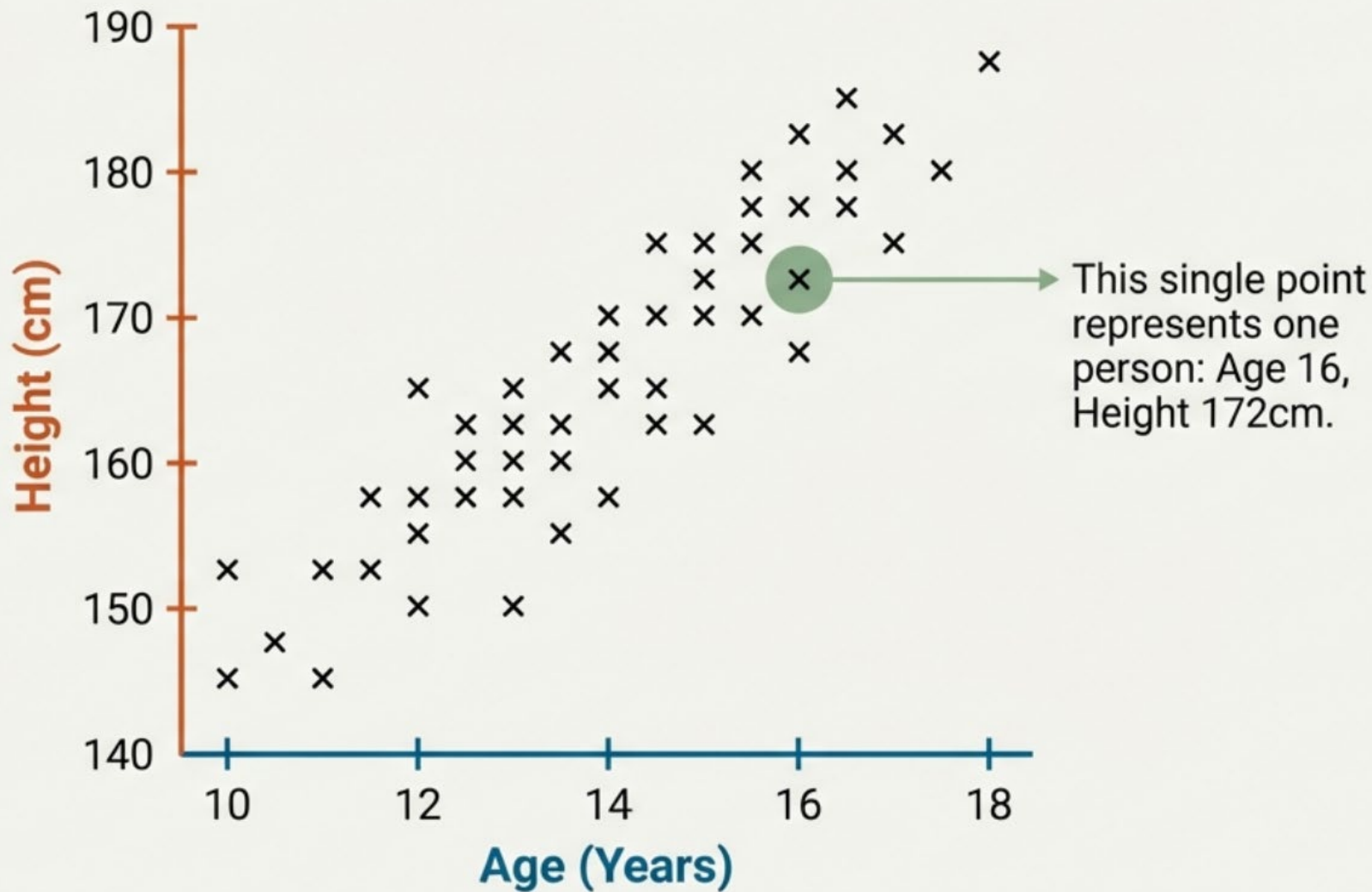
BIVARIATE DATA

Data collected on two variables at the same time.

Height	155	155	157	159	162	167	170	157	154	165
Age	12	13	13	14	15	16	17	14	13	16
Height	148	152	160	166	164	170	172	175	179	183
Age	11	12	13	14	14	15	15	16	17	18
Height	169	171	163	148	161	146	152	164	165	172
Age	17	18	15	12	14	11	13	16	15	18
Height	158	167	174	178	150	156	168	185	177	172
Age	13	14	16	17	12	13	15	18	17	16

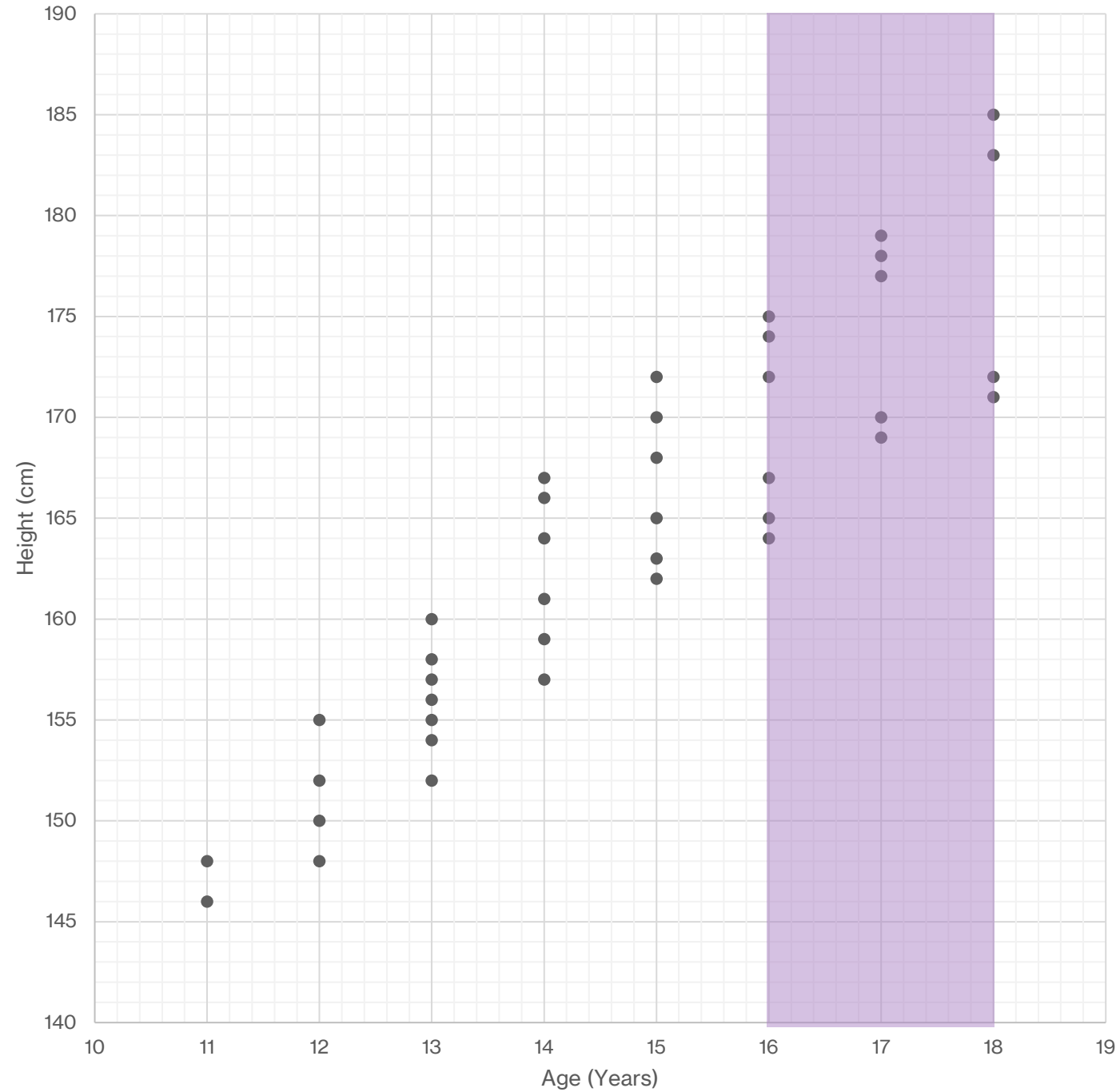
BIVARIATE DATA

We plot one variable against the other to look for relationships.

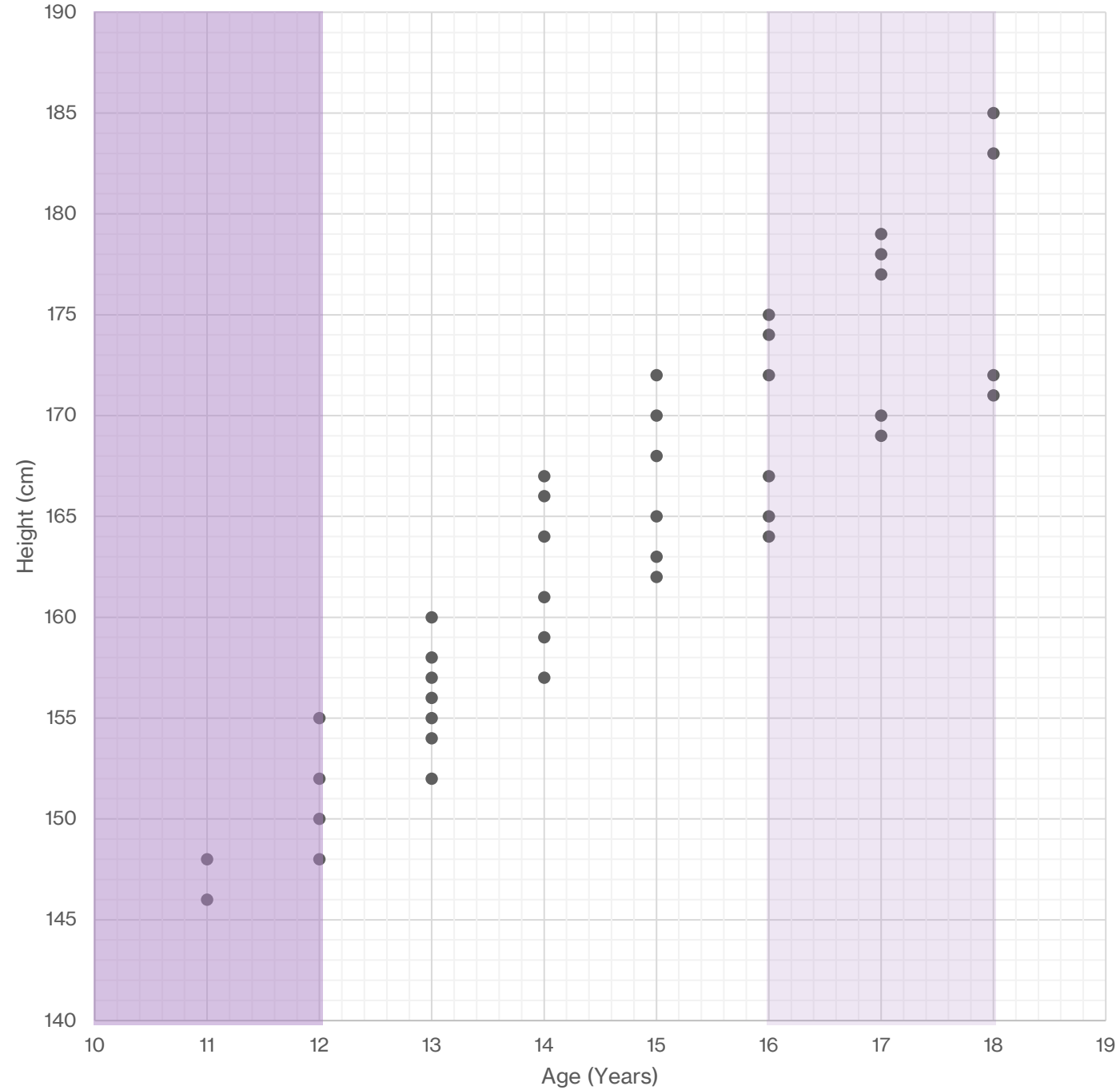


BIVARIATE DATA

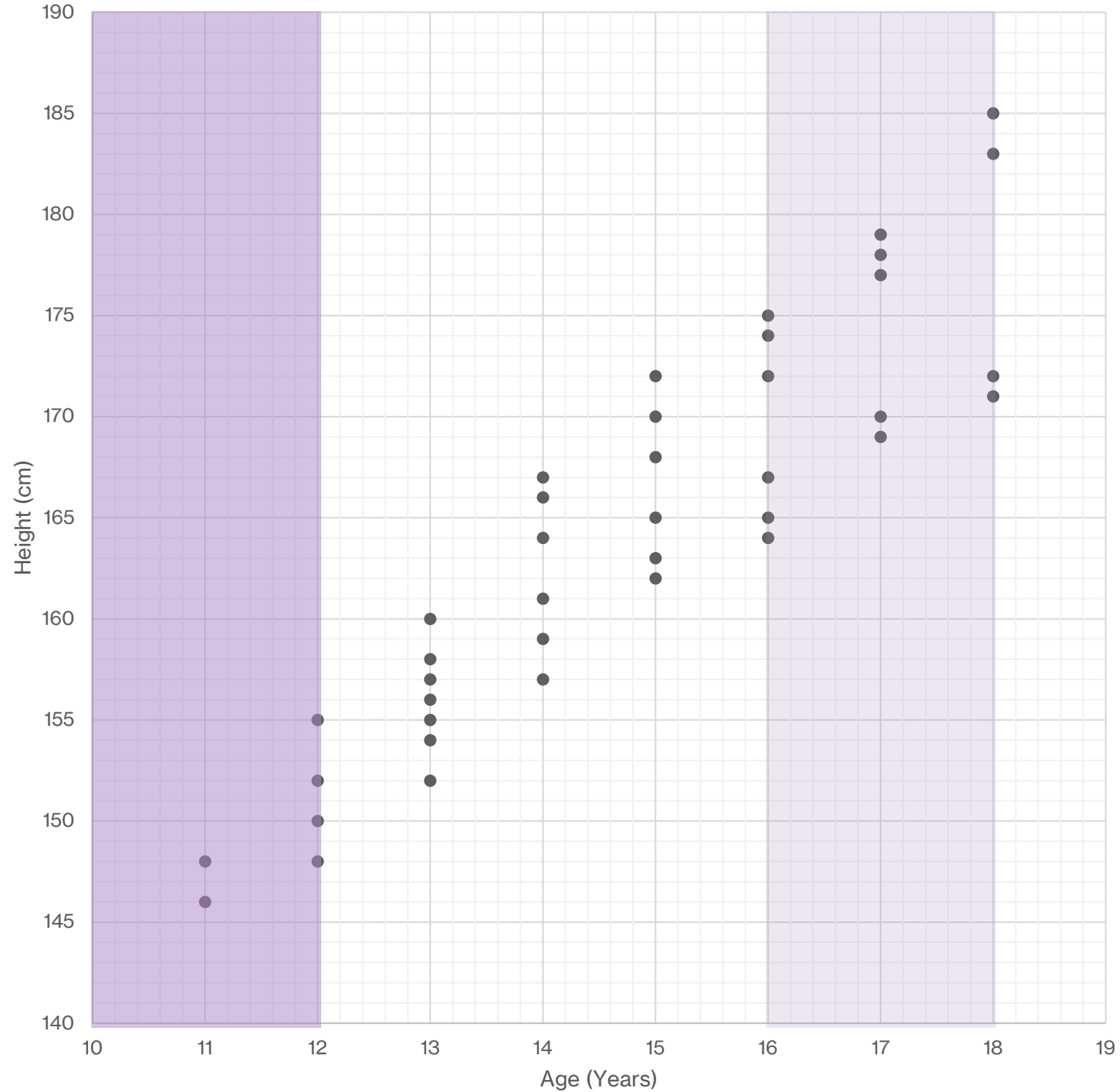
We plot one variable against the other to look for relationships.



BIVARIATE
DATA

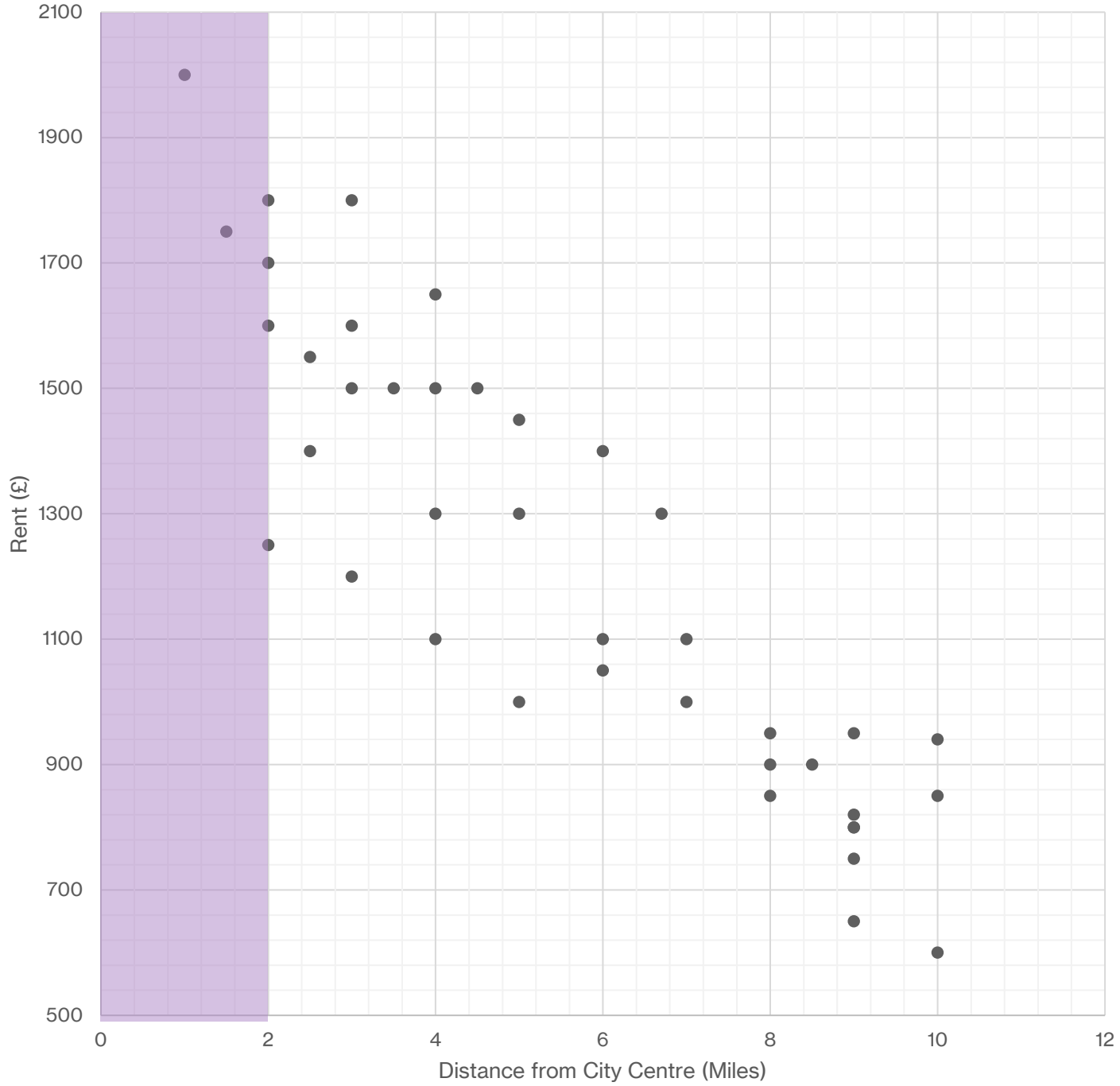


BIVARIATE
DATA

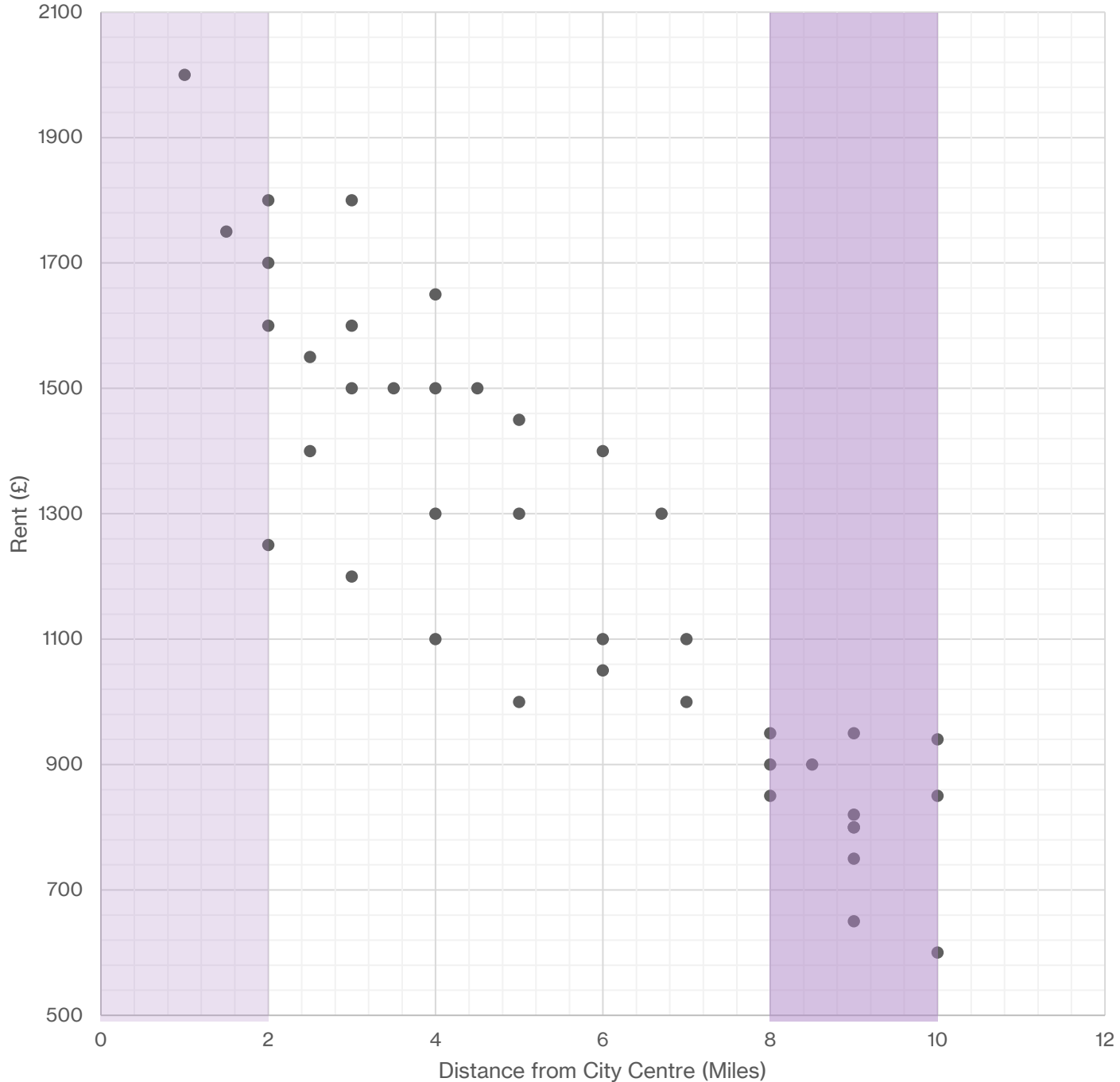


BIVARIATE
DATA

**Positive
Correlation**



BIVARIATE DATA

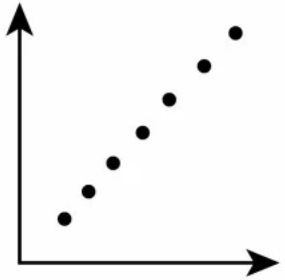


BIVARIATE
DATA

**Negative
Correlation**

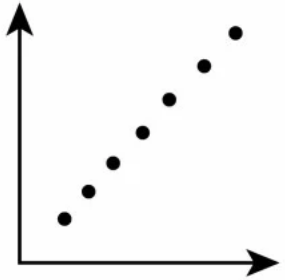
STRENGTH OF CORRELATION

Perfect
Positive



STRENGTH OF CORRELATION

Perfect
Positive

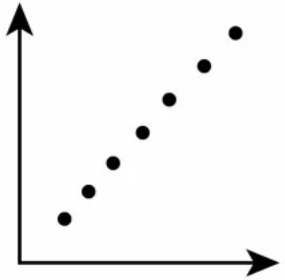


Strong
Positive

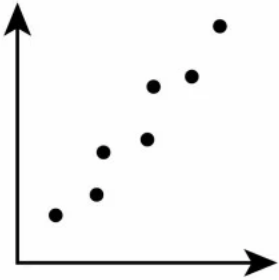


STRENGTH OF CORRELATION

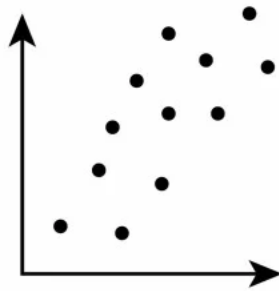
Perfect
Positive



Strong
Positive

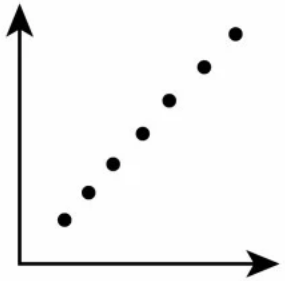


Weak
Positive



STRENGTH OF CORRELATION

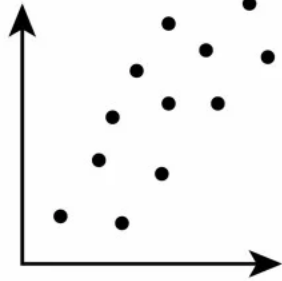
Perfect
Positive



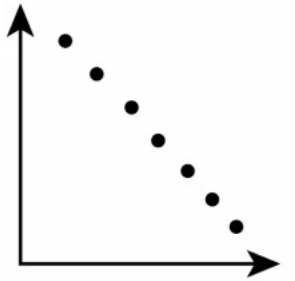
Strong
Positive



Weak
Positive

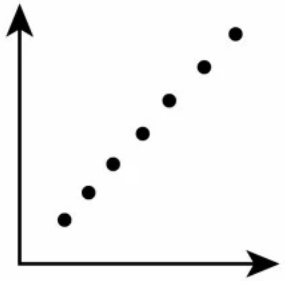


Perfect
Negative



STRENGTH OF CORRELATION

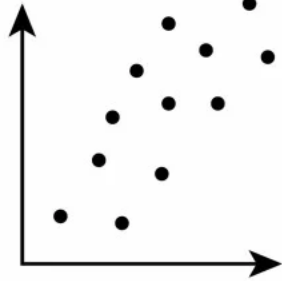
Perfect
Positive



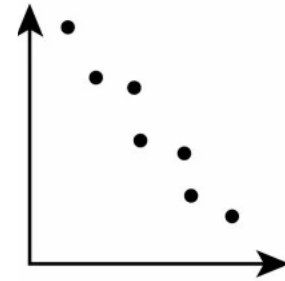
Strong
Positive



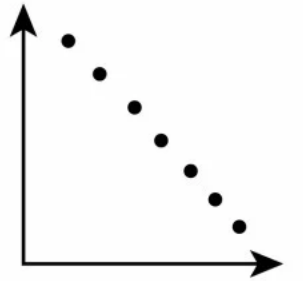
Weak
Positive



strong
Negative

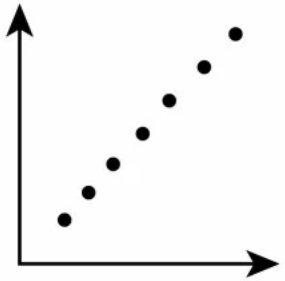


Perfect
Negative



STRENGTH OF CORRELATION

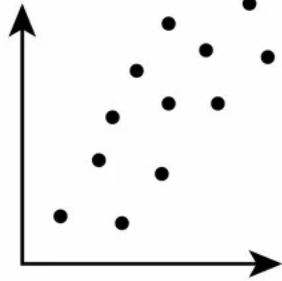
Perfect
Positive



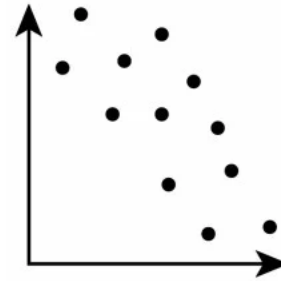
Strong
Positive



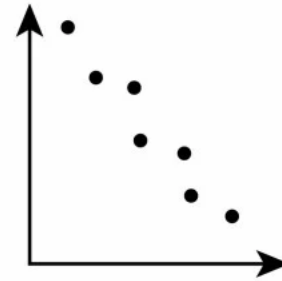
Weak
Positive



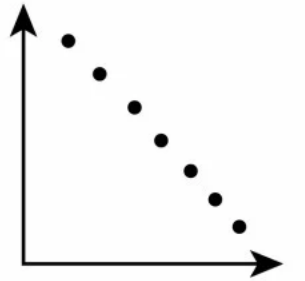
Weak
Negative



strong
Negative

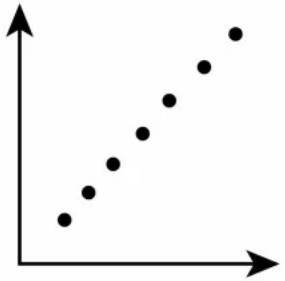


Perfect
Negative



STRENGTH OF CORRELATION

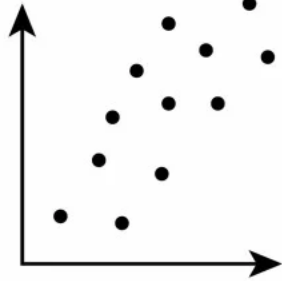
Perfect
Positive



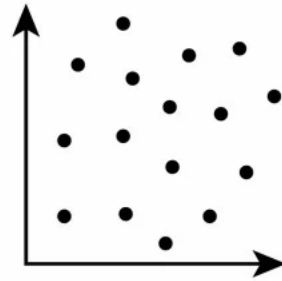
Strong
Positive



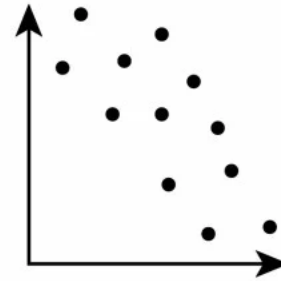
Weak
Positive



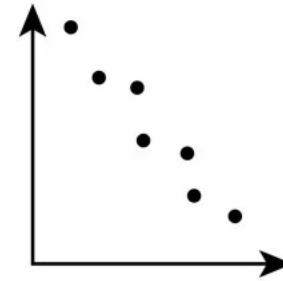
No
Correlation



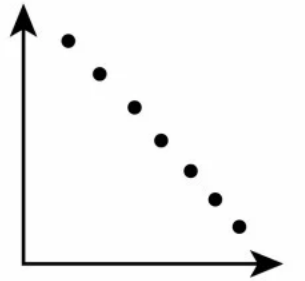
Weak
Negative



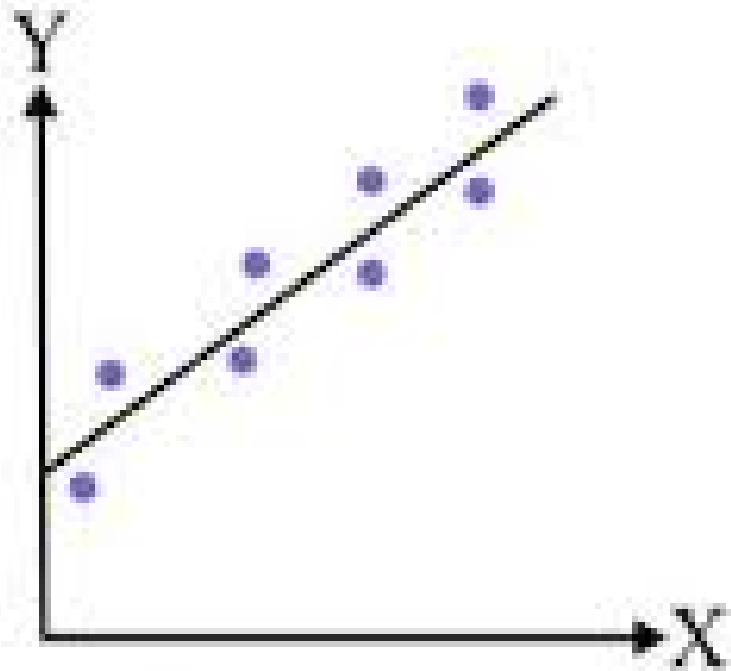
strong
Negative



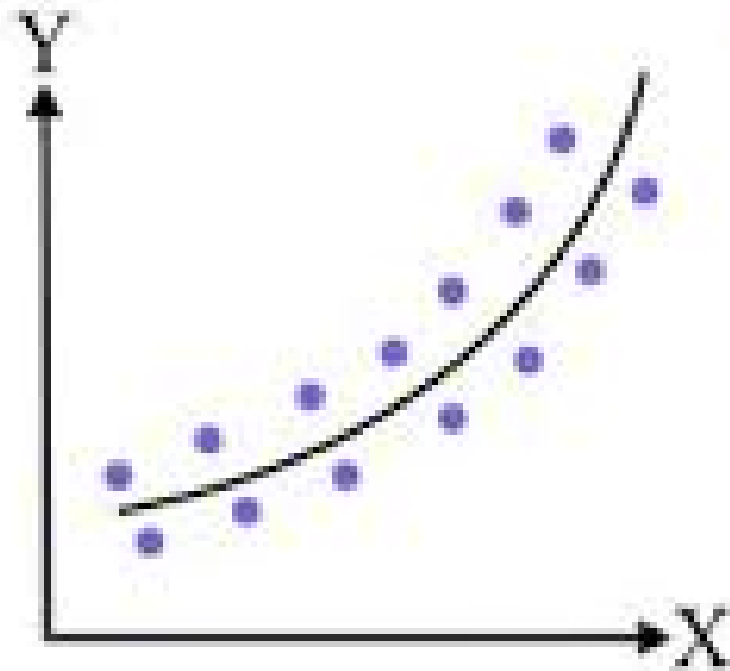
Perfect
Negative



NON-LINEAR CORRELATION



Linear
Correlation



Non-Linear
Correlation

Explanatory Variable

The variable used to predict.
Plotted on the X-Axis.

Explanatory Variable

The variable used to predict.
Plotted on the X-Axis.

Response Variable

The variable being explained.
Plotted on the Y-Axis.

Changes in X cause
changes in Y.

Explanatory Variable

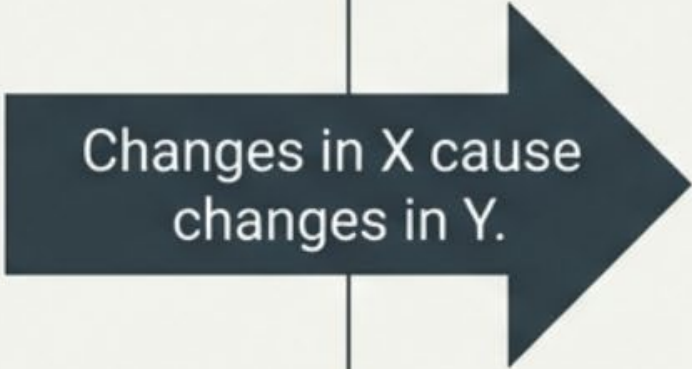
The variable used to predict.
Plotted on the X-Axis.

- Hours of Study

Response Variable

The variable being explained.
Plotted on the Y-Axis.

- Exam Score



Changes in X cause
changes in Y.

Explanatory Variable

The variable used to predict.
Plotted on the X-Axis.

- Hours of Study
- Temperature

Changes in X cause
changes in Y.

Response Variable

The variable being explained.
Plotted on the Y-Axis.

- Exam Score
- Ice Cream Sales

Explanatory Variable

The variable used to predict.
Plotted on the X-Axis.

- Hours of Study
- Temperature
- Age of Car

Changes in X cause
changes in Y.

Response Variable

The variable being explained.
Plotted on the Y-Axis.

- Exam Score
- Ice Cream Sales
- Car Value

Explanatory Variable (Independent)



Changes in X cause
changes in Y.

Response Variable (Dependent)



Explanatory Variable (Independent)

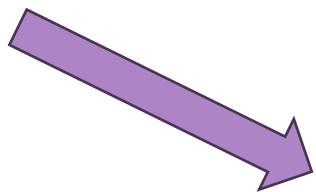


Changes in X cause
changes in Y.

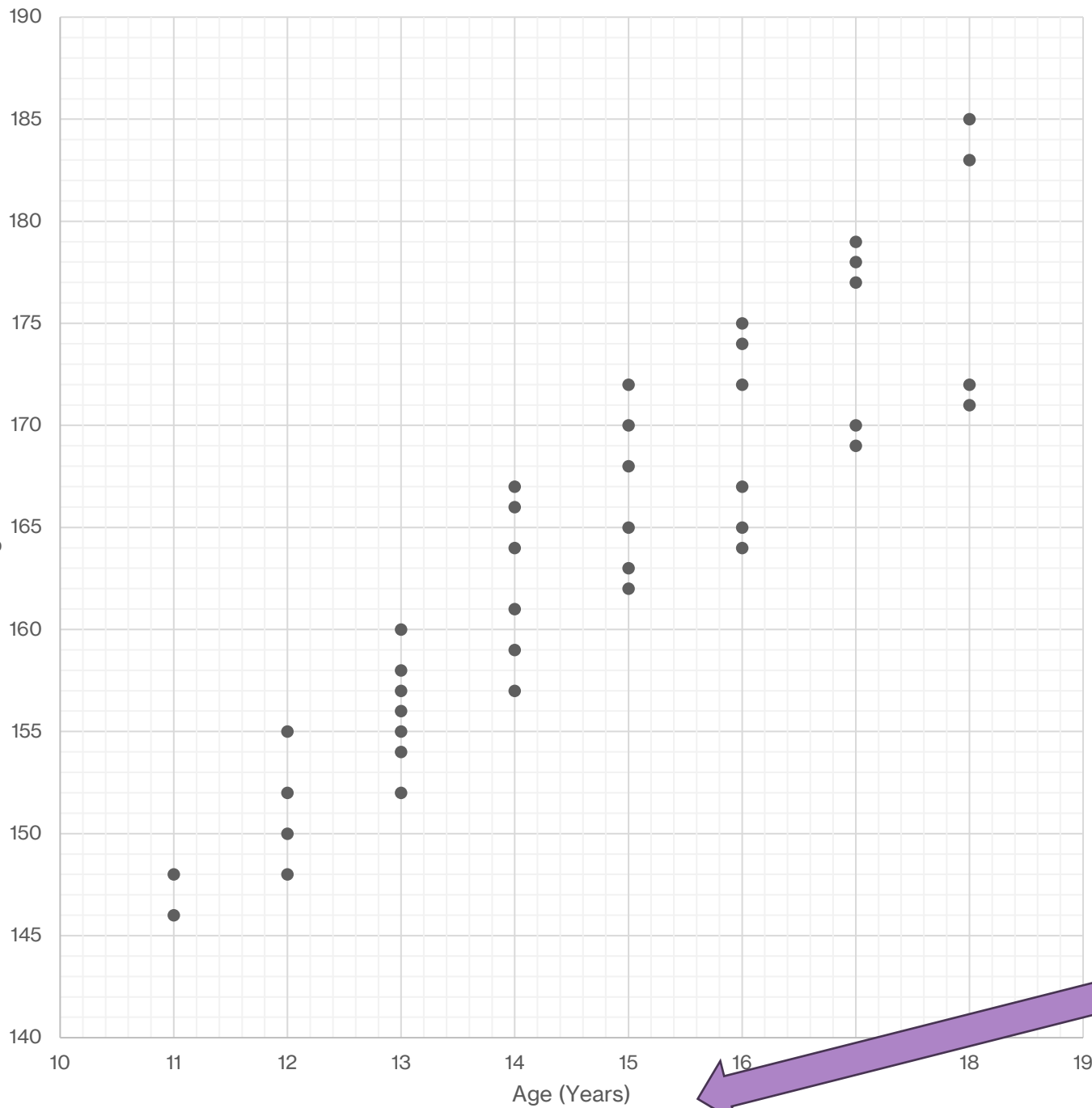
Response Variable (Dependent)



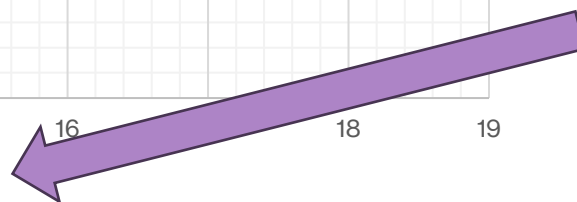
**Response /
Dependent
Variable**



Height (cm)



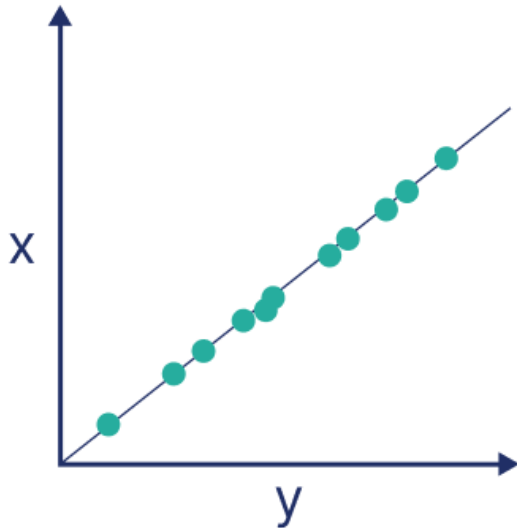
**Explanatory /
Independent
Variable**



PEARSON'S CORRELATION CO-EFFICIENT

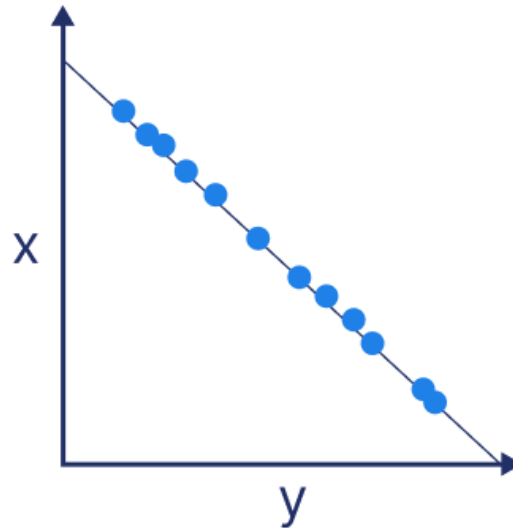
Perfect positive correlation

$$r = 1$$



Perfect negative correlation

$$r = -1$$

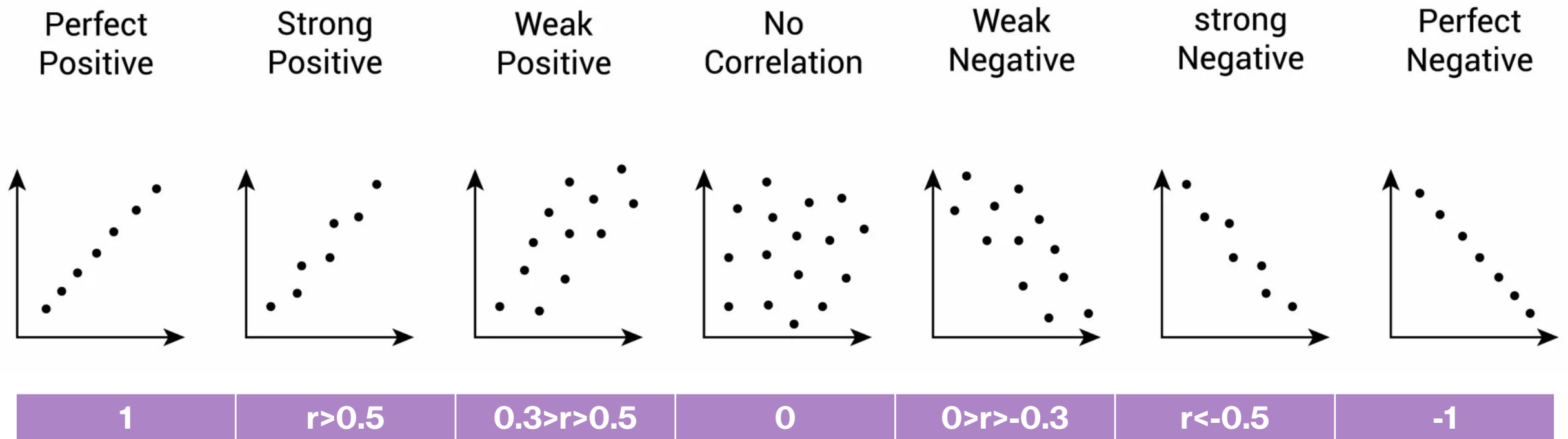


- The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation.
- It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.

PEARSON'S CORRELATION CO-EFFICIENT

r value	Strength	Direction
Greater than 0.5	Strong	+ve
Between 0.3 and 0.5	Moderate	+ve
Between 0 and 0.3	Weak	+ve
0	None	None
Between 0 and -0.3	Weak	-ve
Between -0.3 and -0.5	Moderate	-ve
Less than -0.5	Strong	-ve

STRENGTH OF CORRELATION



PEARSON'S CORRELATION CO-EFFICIENT

The Pearson correlation coefficient is a good choice when all the following are true:

- **Both variables are quantitative**

PEARSON'S CORRELATION CO-EFFICIENT

The Pearson correlation coefficient is a good choice when all the following are true:

- **Both variables are quantitative**
- **The variables are normally distributed**

PEARSON'S CORRELATION CO-EFFICIENT

The Pearson correlation coefficient is a good choice when all the following are true:

- **Both variables are quantitative**
- **The variables are normally distributed**
- **The data has no outliers**

PEARSON'S CORRELATION CO-EFFICIENT

The Pearson correlation coefficient is a good choice when all the following are true:

- **Both variables are quantitative**
- **The variables are normally distributed**
- **The data has no outliers**
- **The relationship is linear**

CALCULATING PEARSON'S CORRELATION COEFFICIENT

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

CALCULATING
PEARSON'S
CORRELATION
CO-EFFICIENT

Revision (Hours)	Test Score (%)
0	68
2	69
4	74
4	82
5	77
6	84
8	85
10	99

CALCULATING
PEARSON'S
CORRELATION
CO-EFFICIENT

x	y
0	68
2	69
4	74
4	82
5	77
6	84
8	85
10	99
$\Sigma x=39$	$\Sigma y=638$

CALCULATING
PEARSON'S
CORRELATION
CO-EFFICIENT

x	x ²	y	y ²
0	0	68	4624
2	4	69	4761
4	16	74	5476
4	16	82	6724
5	25	77	5929
6	36	84	7056
8	64	85	7225
10	100	99	9801
Σx=39	Σx ² =261	Σy=638	Σy ² =51596

CALCULATING
PEARSON'S
CORRELATION
CO-EFFICIENT

x	y	xy
0	68	0
2	69	138
4	74	296
4	82	328
5	77	385
6	84	504
8	85	680
10	99	990
$\Sigma x=39$	$\Sigma y=638$	$\Sigma xy=3321$

CALCULATING PEARSON'S CORRELATION CO-EFFICIENT

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

$$n = 8$$

$$\sum x = 39$$

$$\sum y = 638$$

$$\sum x^2 = 261$$

$$\sum y^2 = 51596$$

$$\sum xy = 3321$$

CALCULATING PEARSON'S CORRELATION CO-EFFICIENT

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

$$n = 8$$

$$\sum x = 39$$

$$\sum y = 638$$

$$\sum x^2 = 261$$

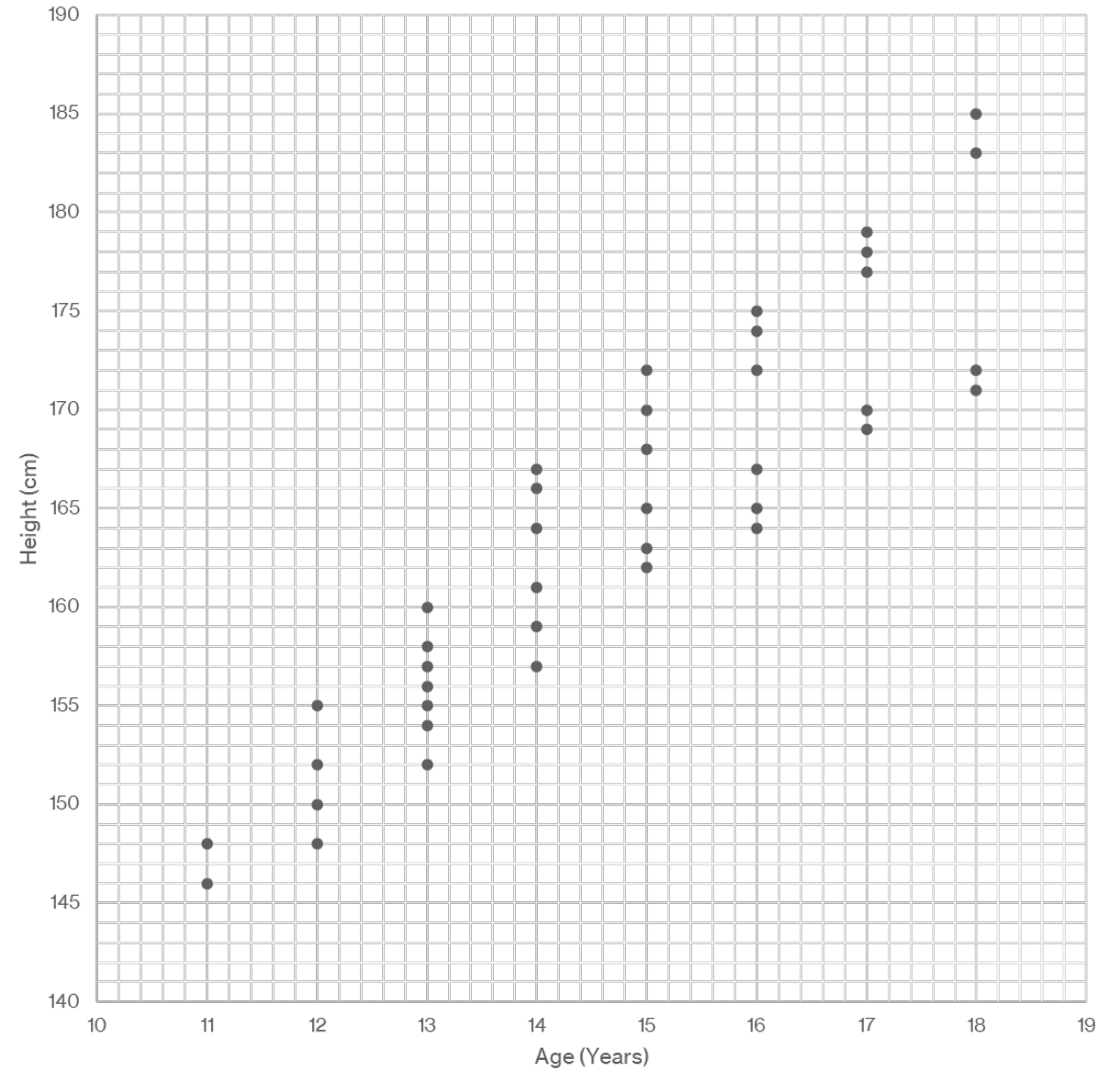
$$\sum y^2 = 51596$$

$$\sum xy = 3321$$

$$r = \frac{(8 \times 3321) - (39 \times 638)}{\sqrt{[(8 \times 261) - 39^2][(8 \times 51596) - 638^2]}}$$

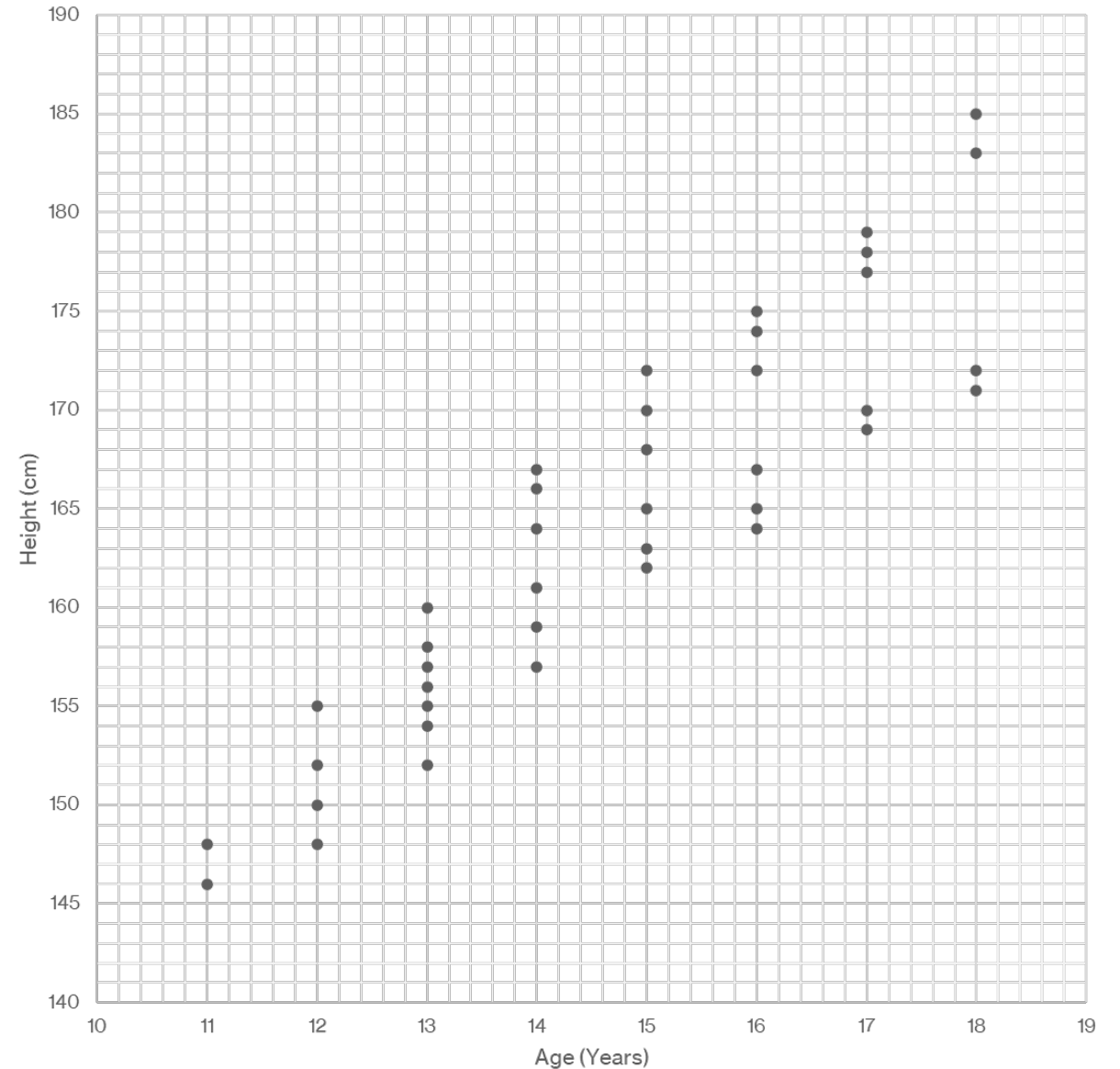
$$r = 0.935876$$

CORRELATION VS CAUSATION

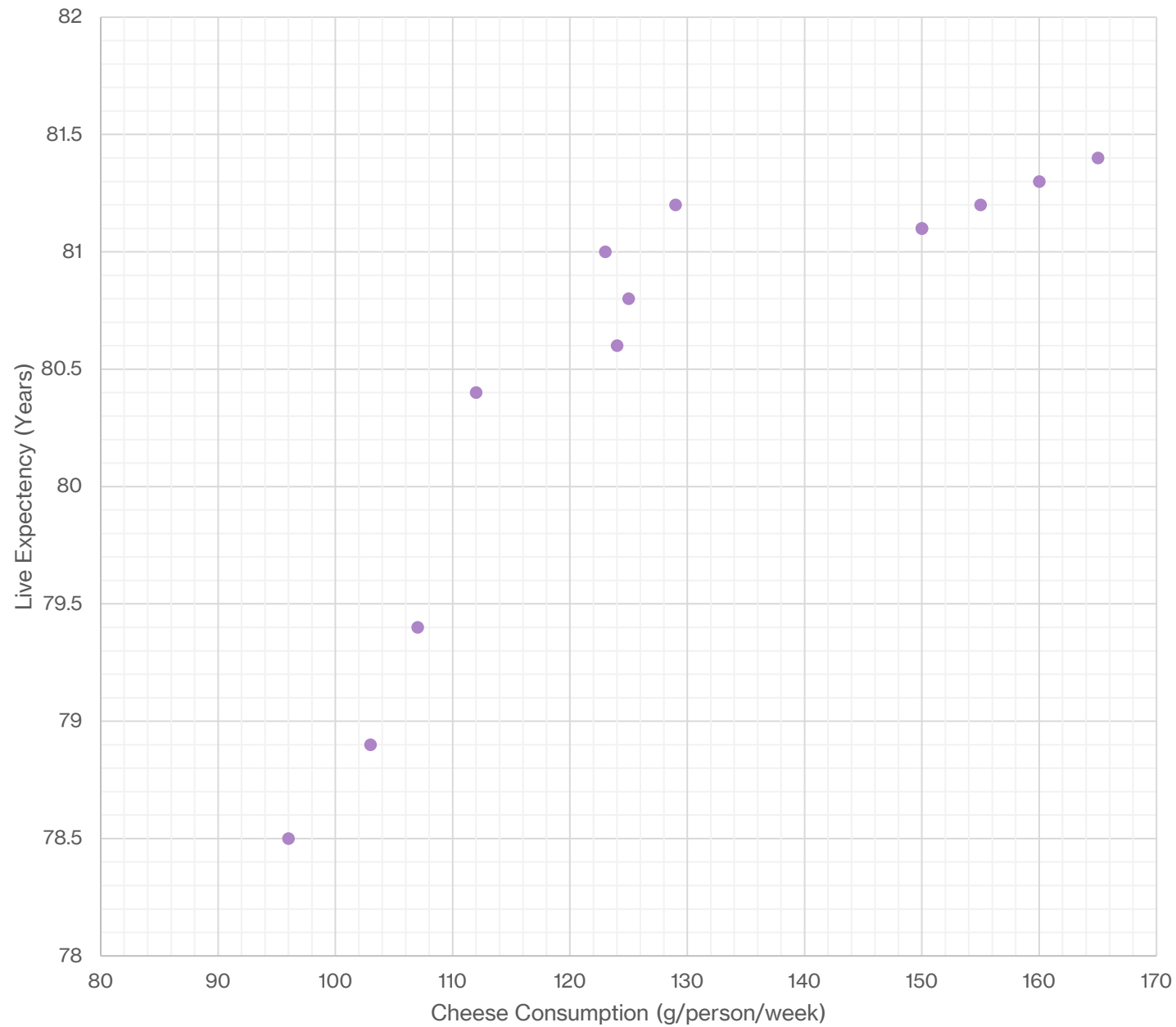


CORRELATION VS CAUSATION

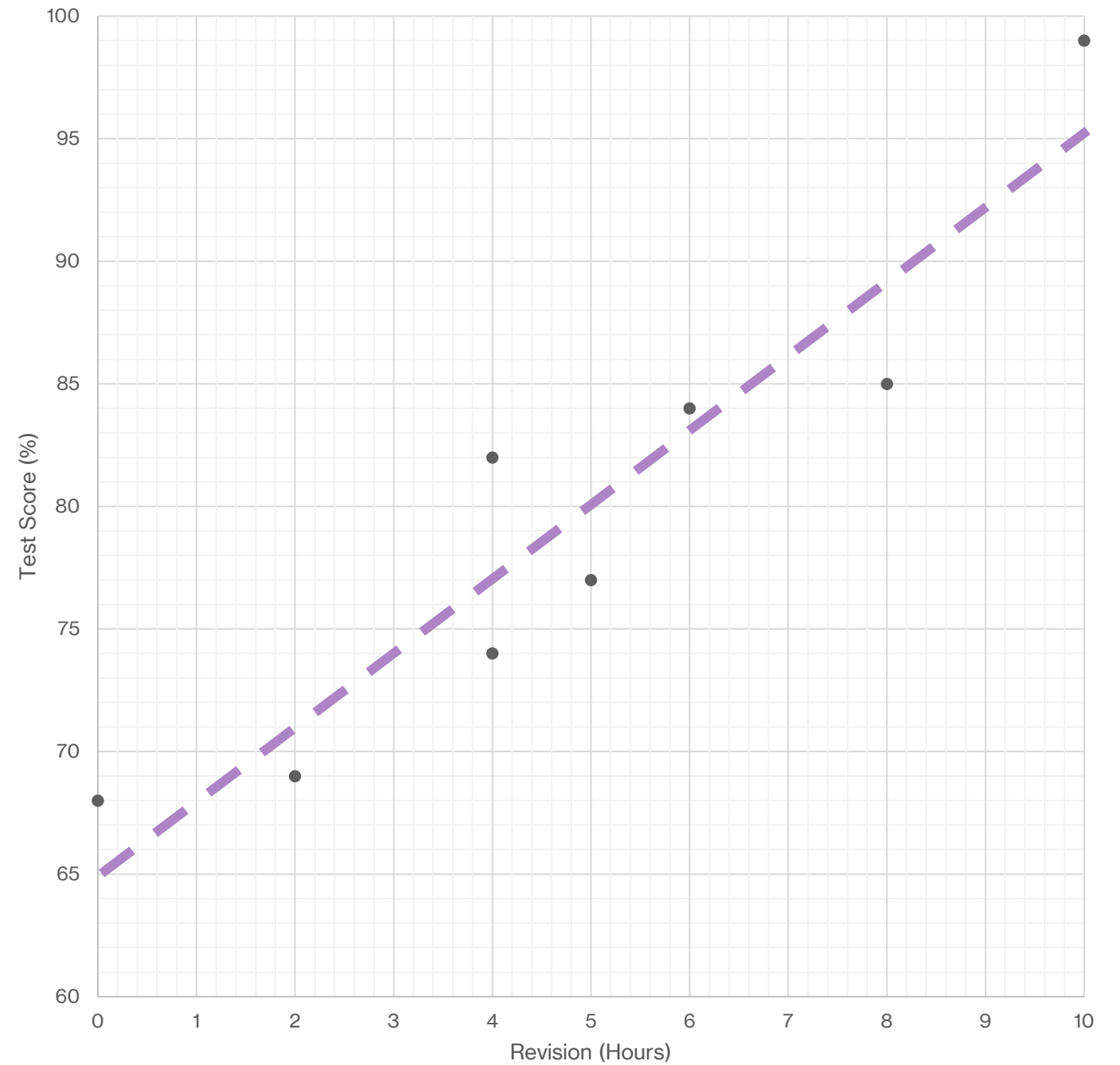
**Age causes a
change in height**



CORRELATION $\neq$
CAUSATION



LINE OF BEST FIT

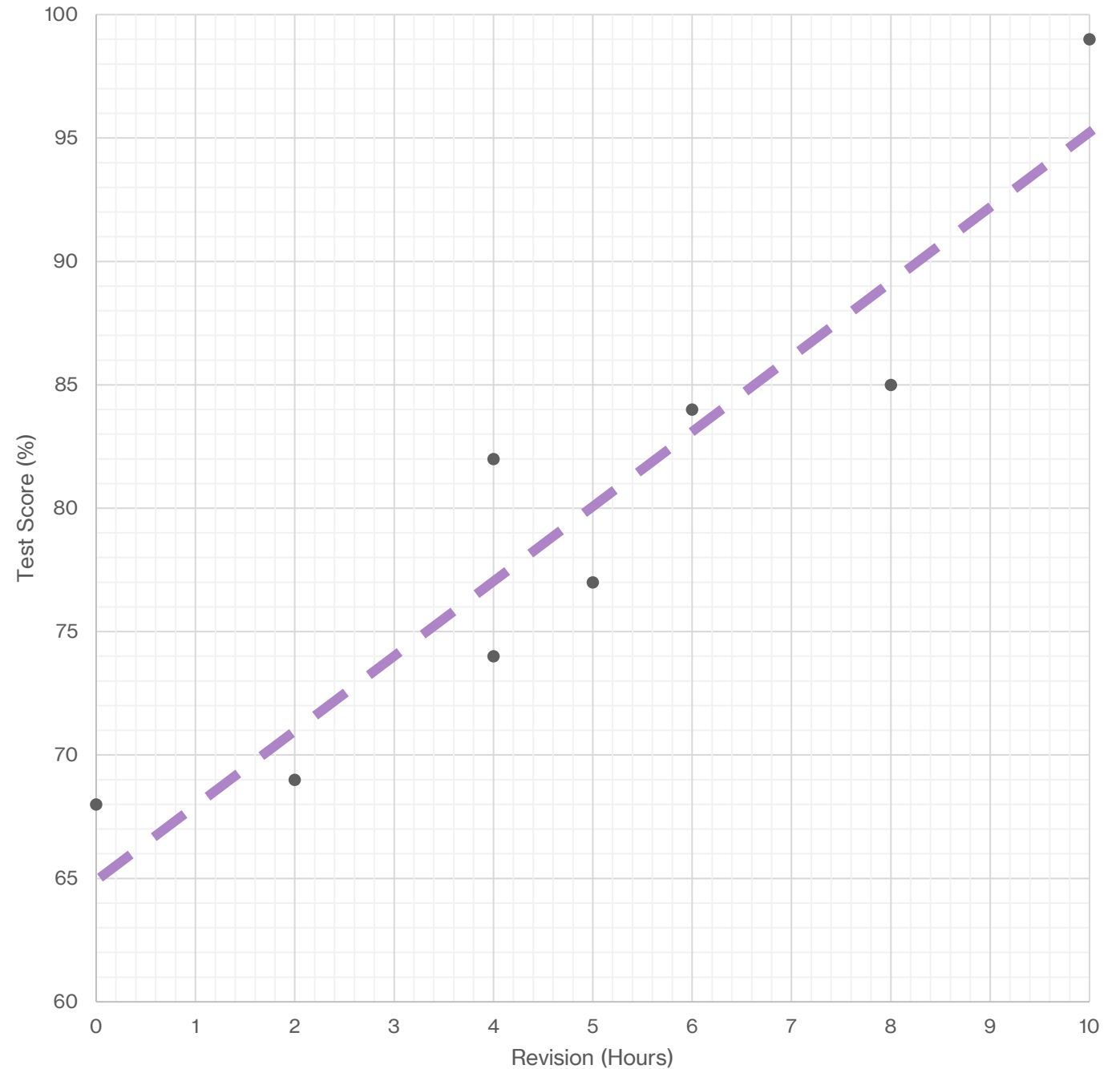


LINE OF BEST FIT

Least Squares Regression Line

y on x regression line

x on y regression line

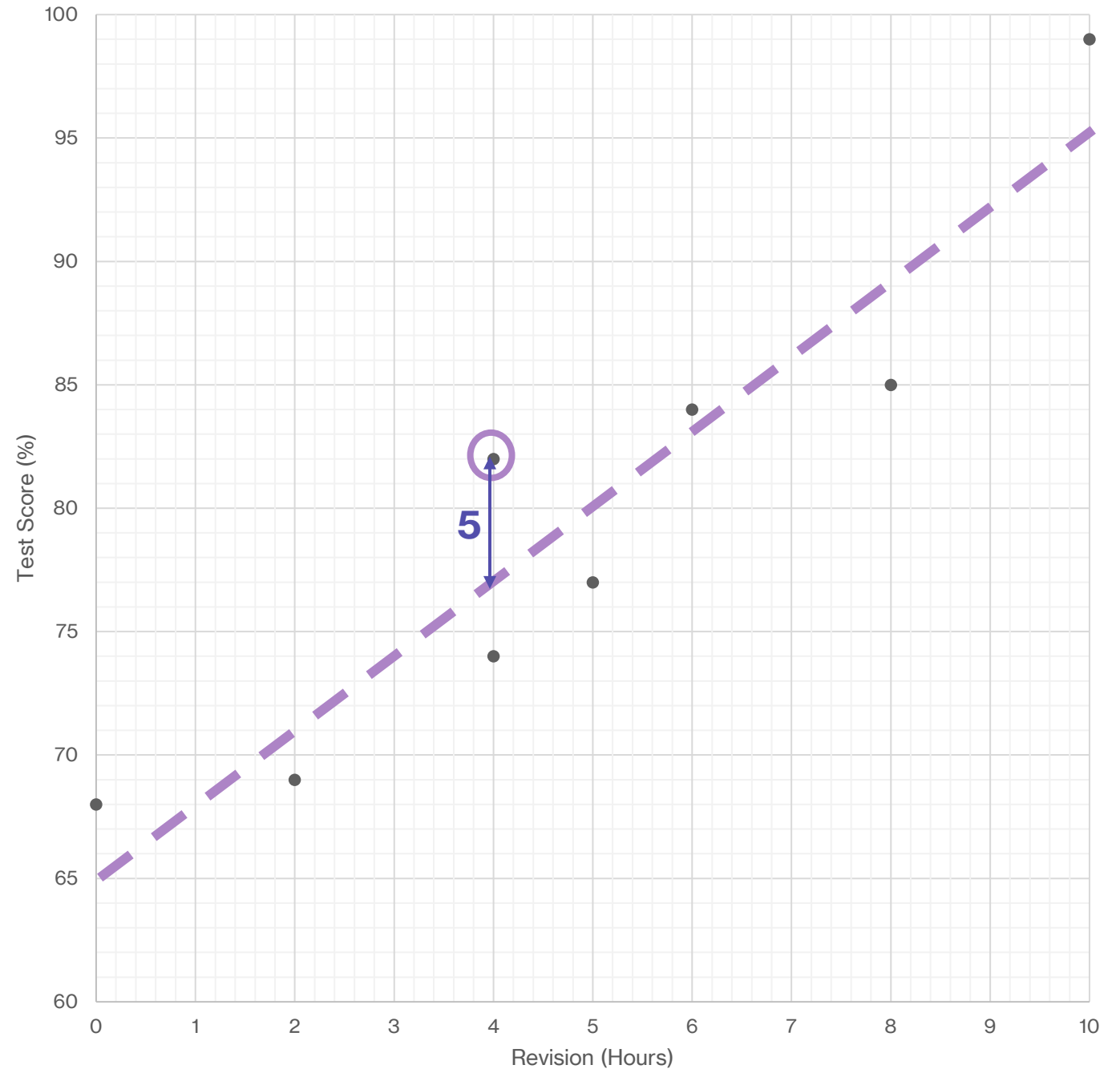


LINE OF BEST FIT

Least Squares Regression Line

y on x regression line

~~x on y regression line~~

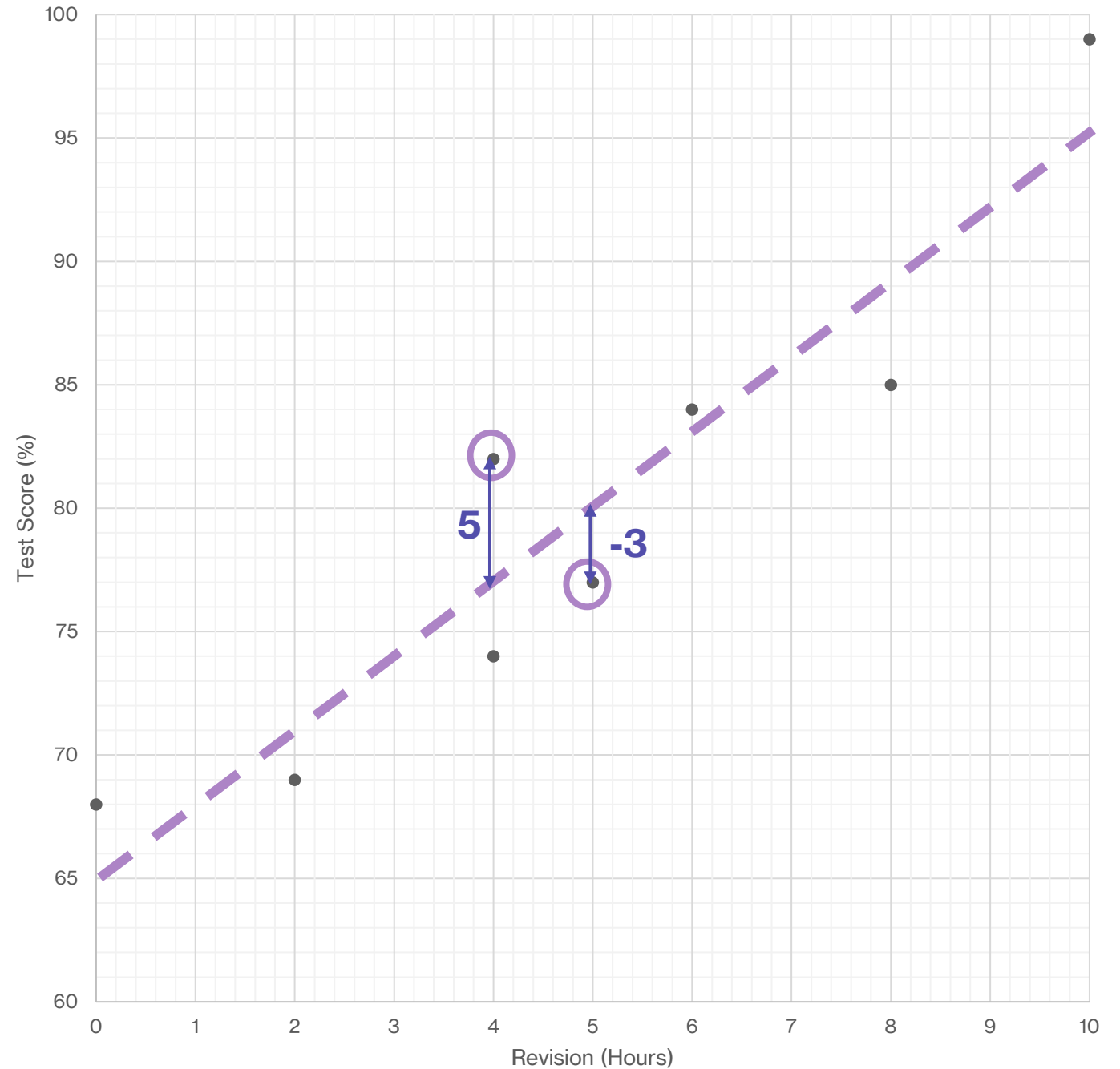


LINE OF BEST FIT

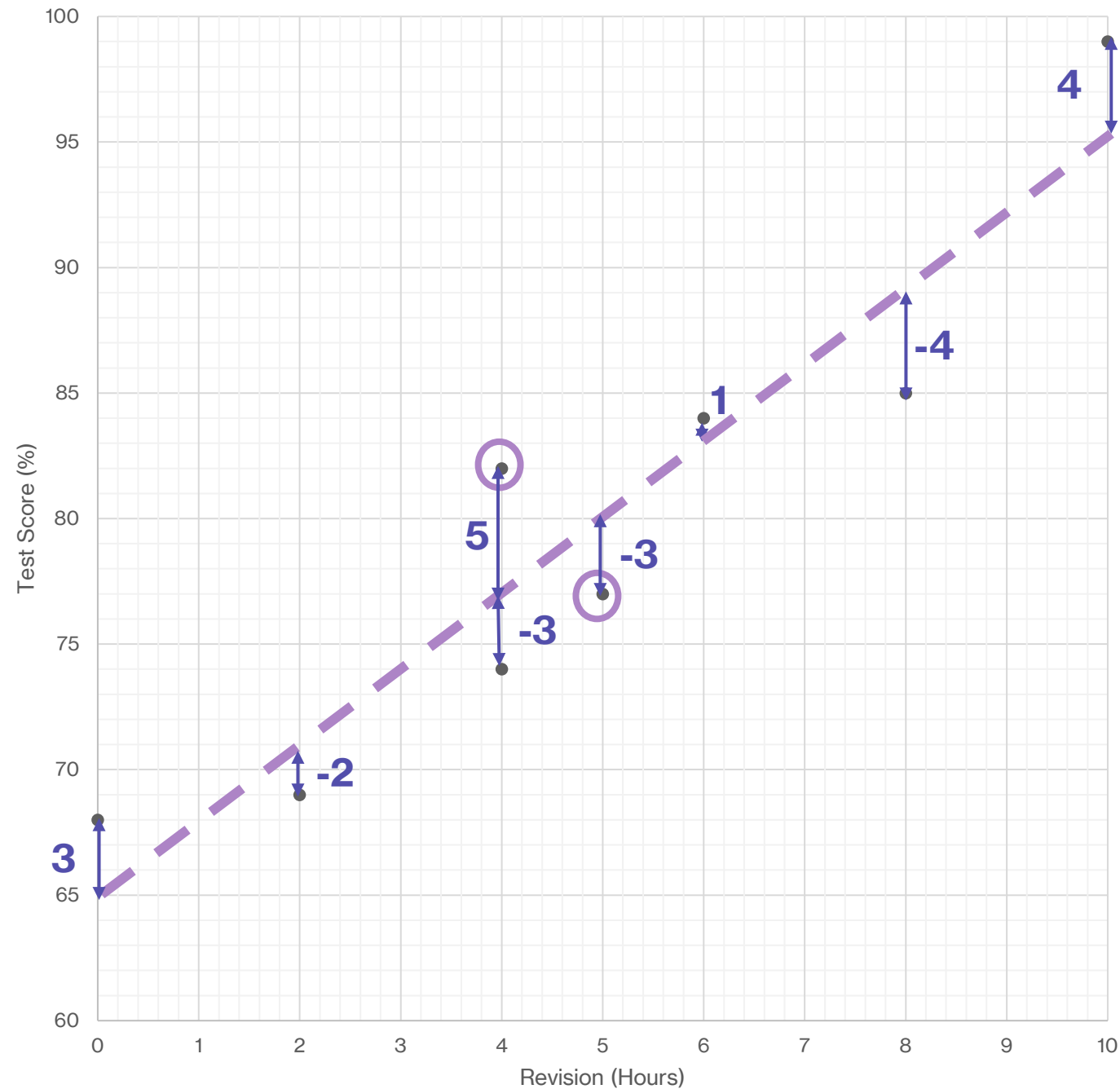
Least Squares Regression Line

y on x regression line

~~x on y regression line~~

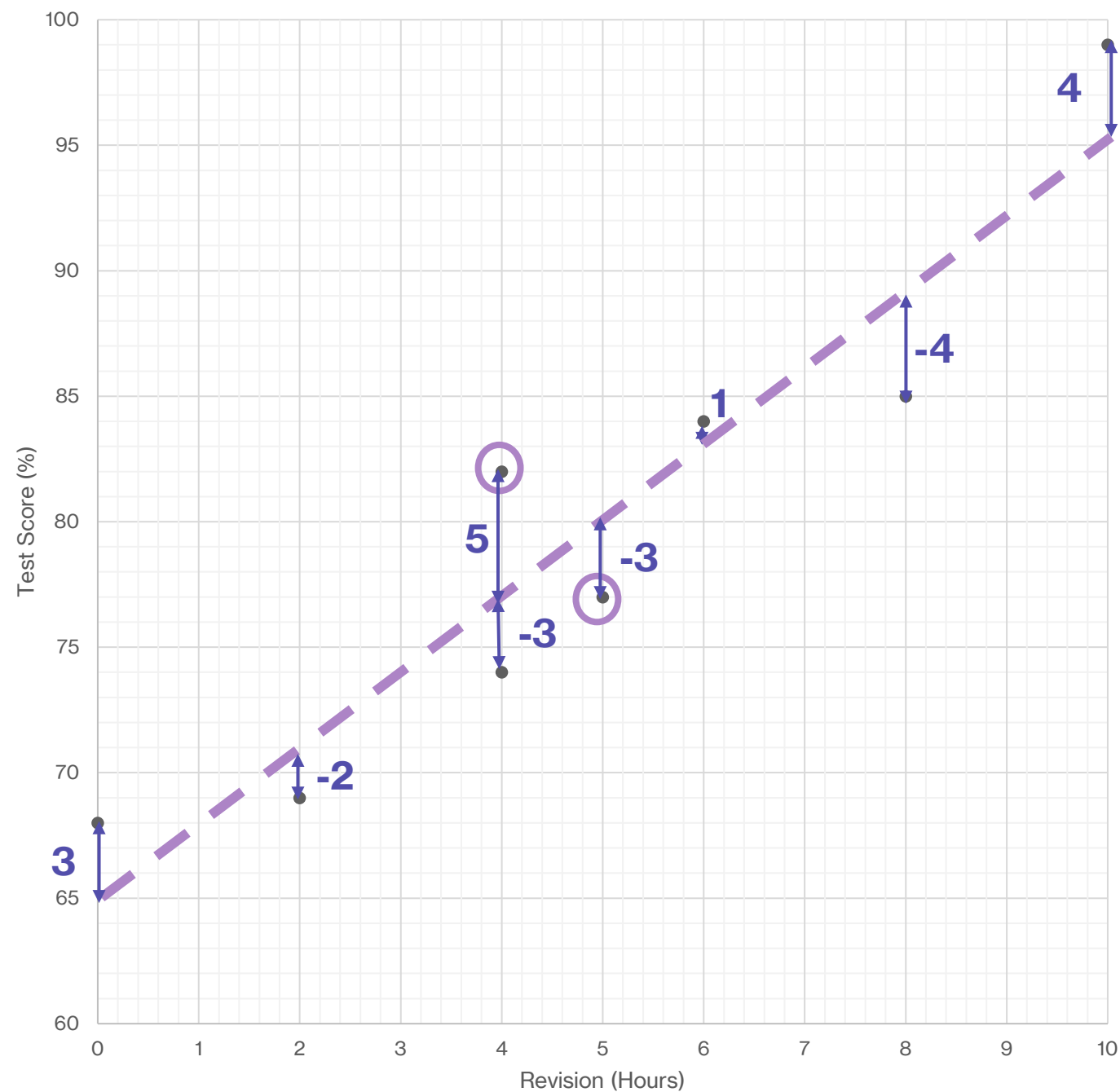


LINE OF BEST FIT



LINE OF BEST FIT

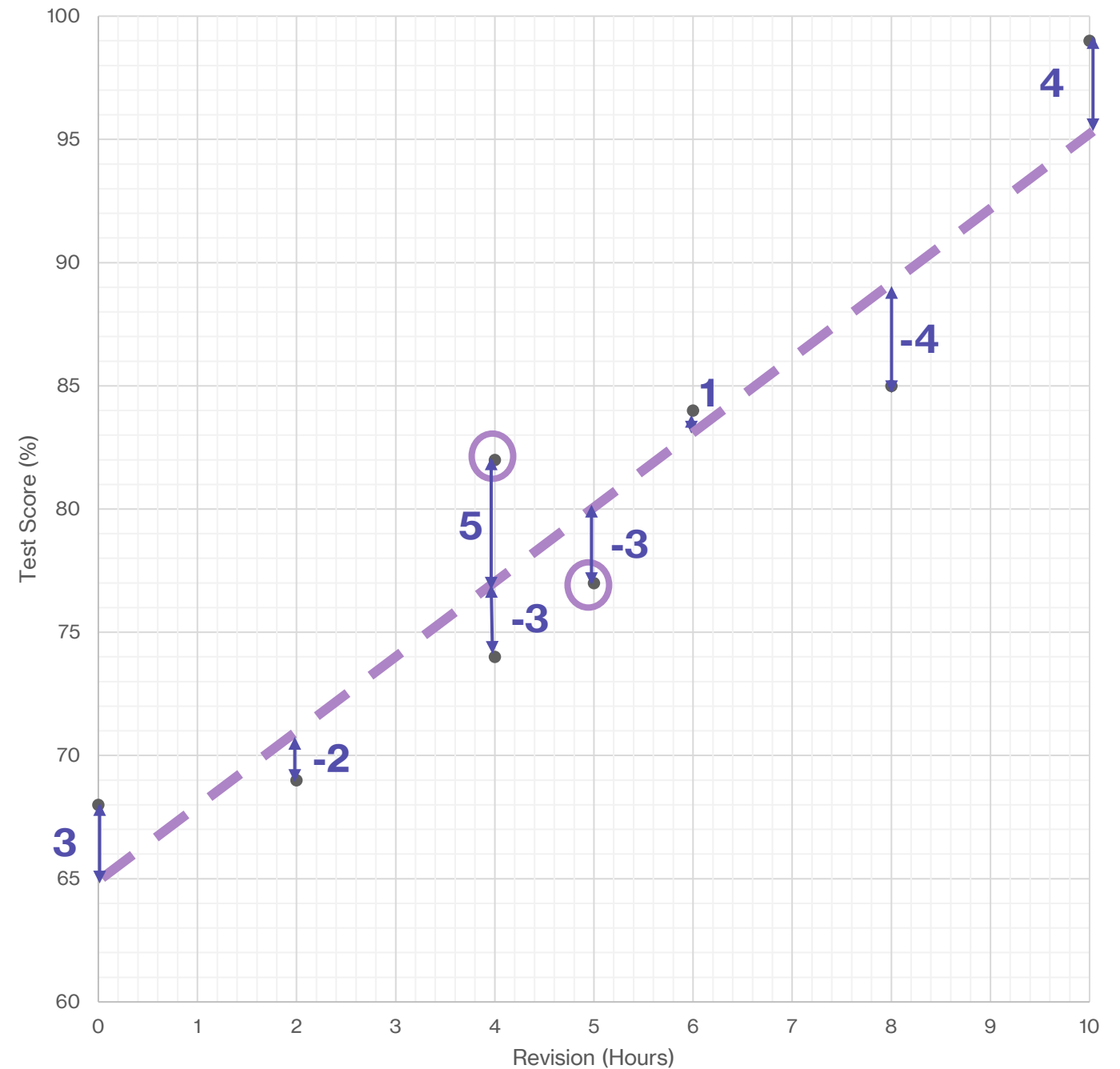
3	-2	5	-3	-3	1	-4	4
9	4	25	9	9	1	16	16



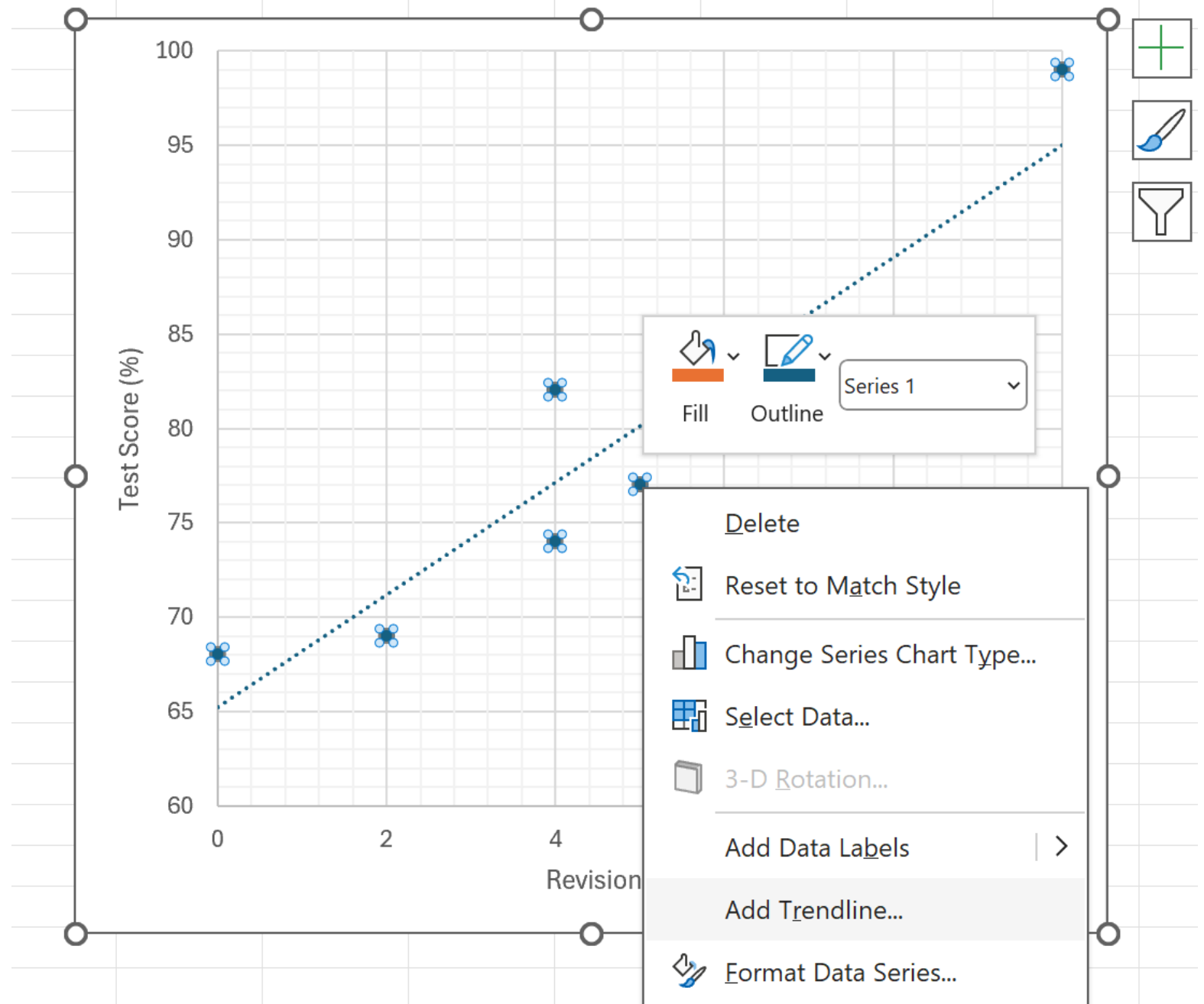
LINE OF BEST FIT

3	-2	5	-3	-3	1	-4	4
9	4	25	9	9	1	16	16

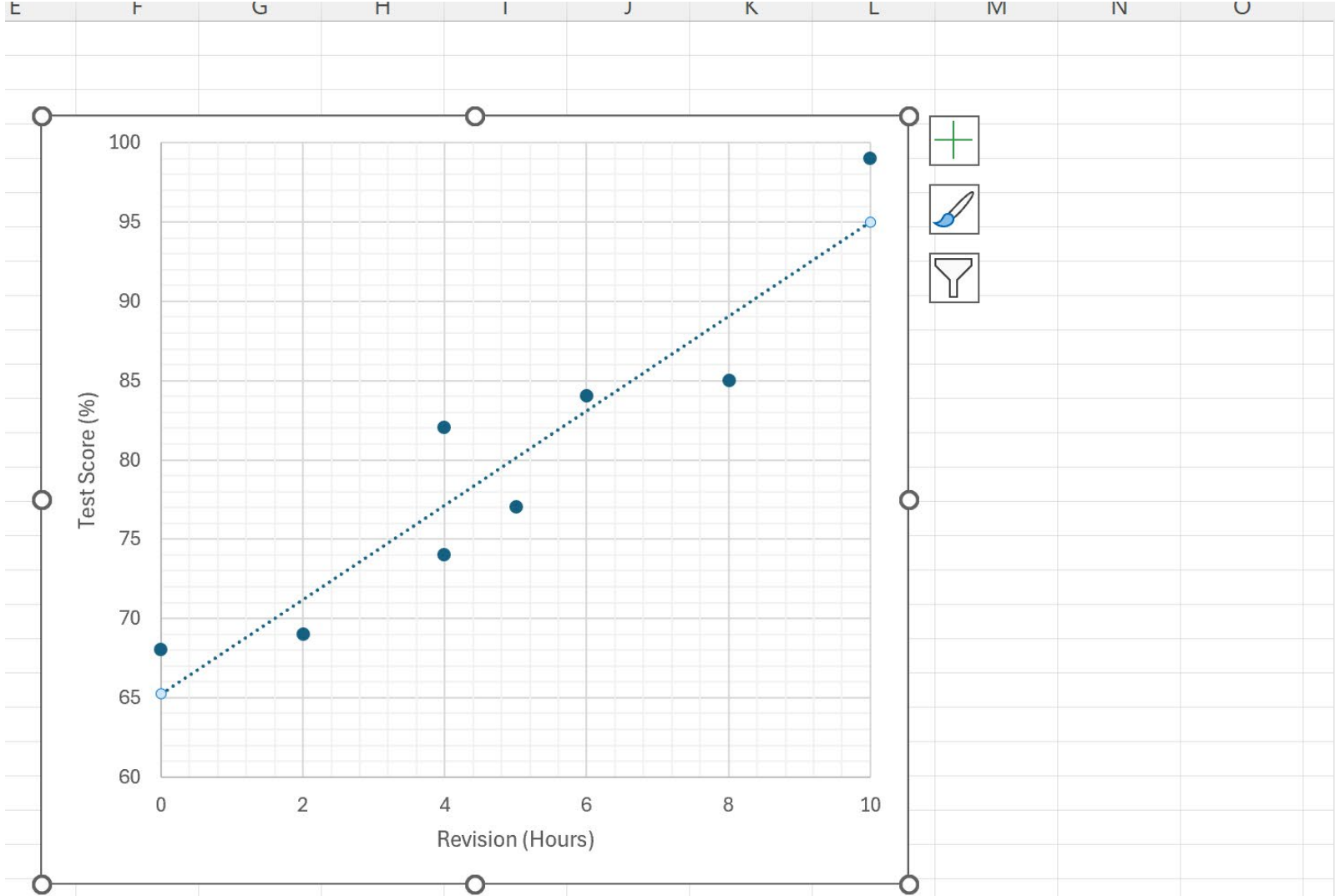
$$\Sigma = 89$$



LINE OF BEST FIT



LINE OF BEST FIT



Format Trendline

Trendline Options

☐ Exponential

☒ Linear

☐ Logarithmic

☐ Polynomial

Order

☐ Power

☐ Moving Average

Period

Trendline Name

☒ Automatic

☐ Custom

Linear (Series1)

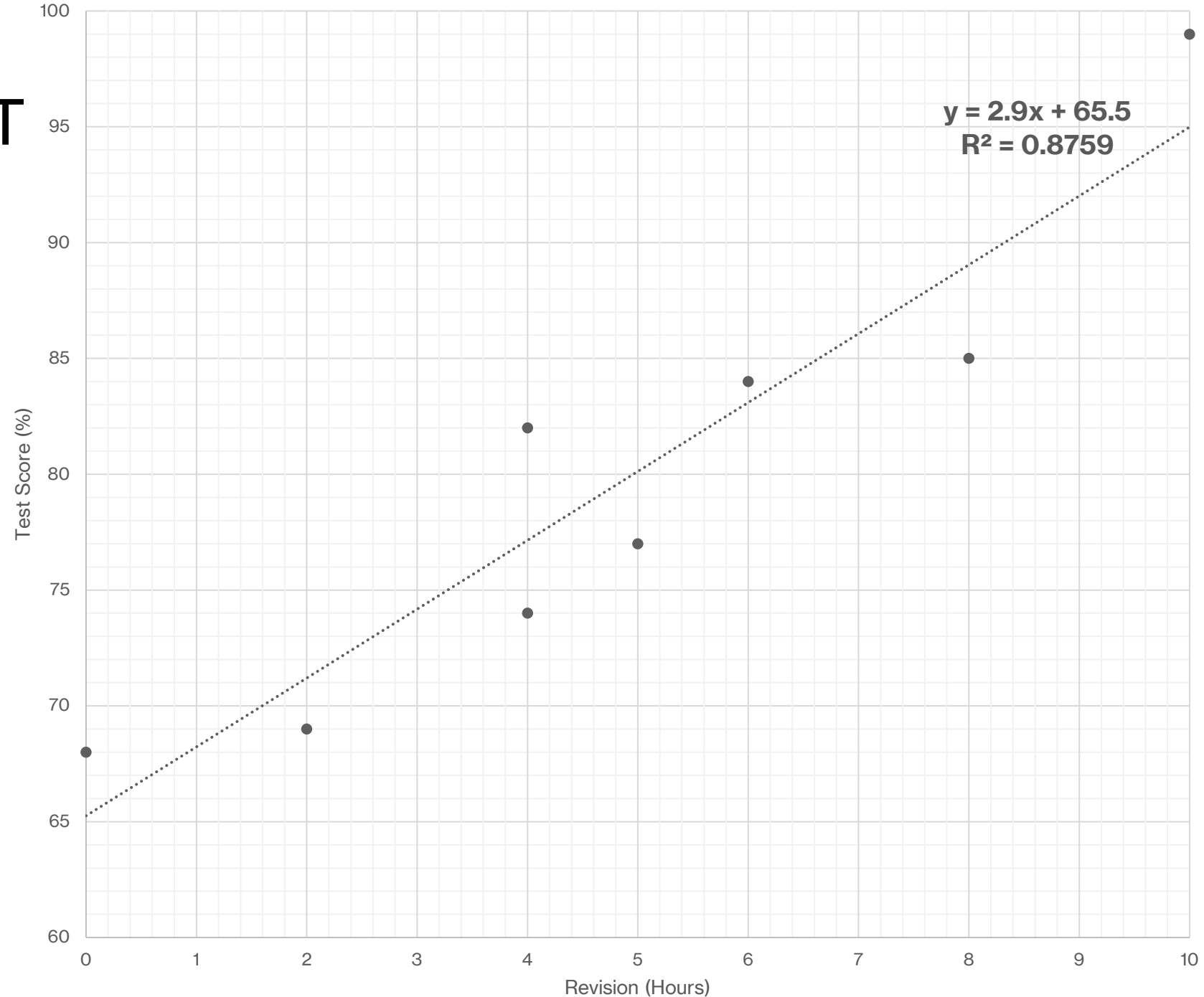
Forecast

Forward

Backward

☐ Set Intercept

LINE OF BEST FIT



Response Variable (Orange)

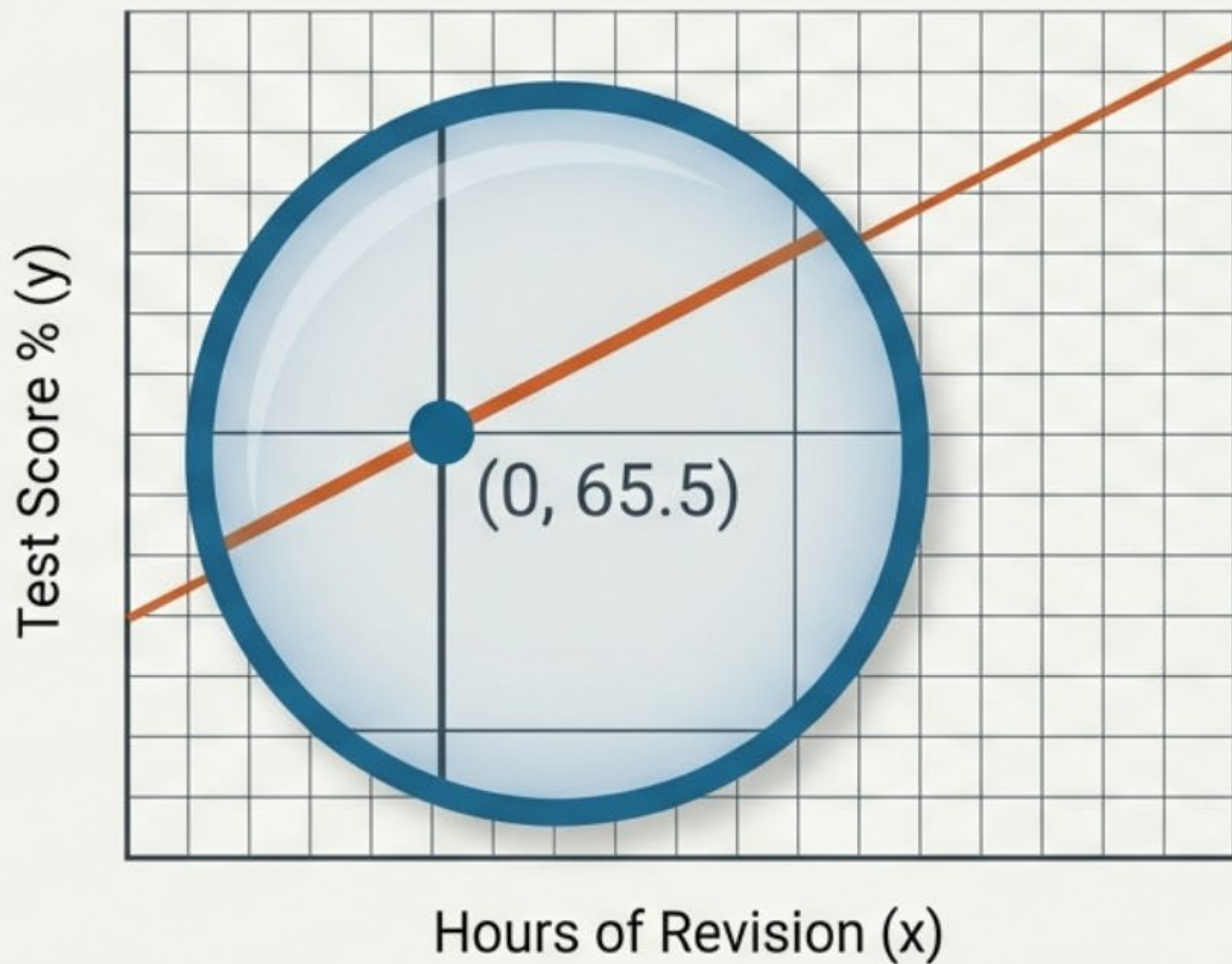
Explanatory Variable (Blue)

$$y = a + bx$$

The Intercept (Constant).
Value of y when $x=0$.

The Regression Coefficient (Gradient).
The Rate of Change.

Interpreting 'a': The Starting Point

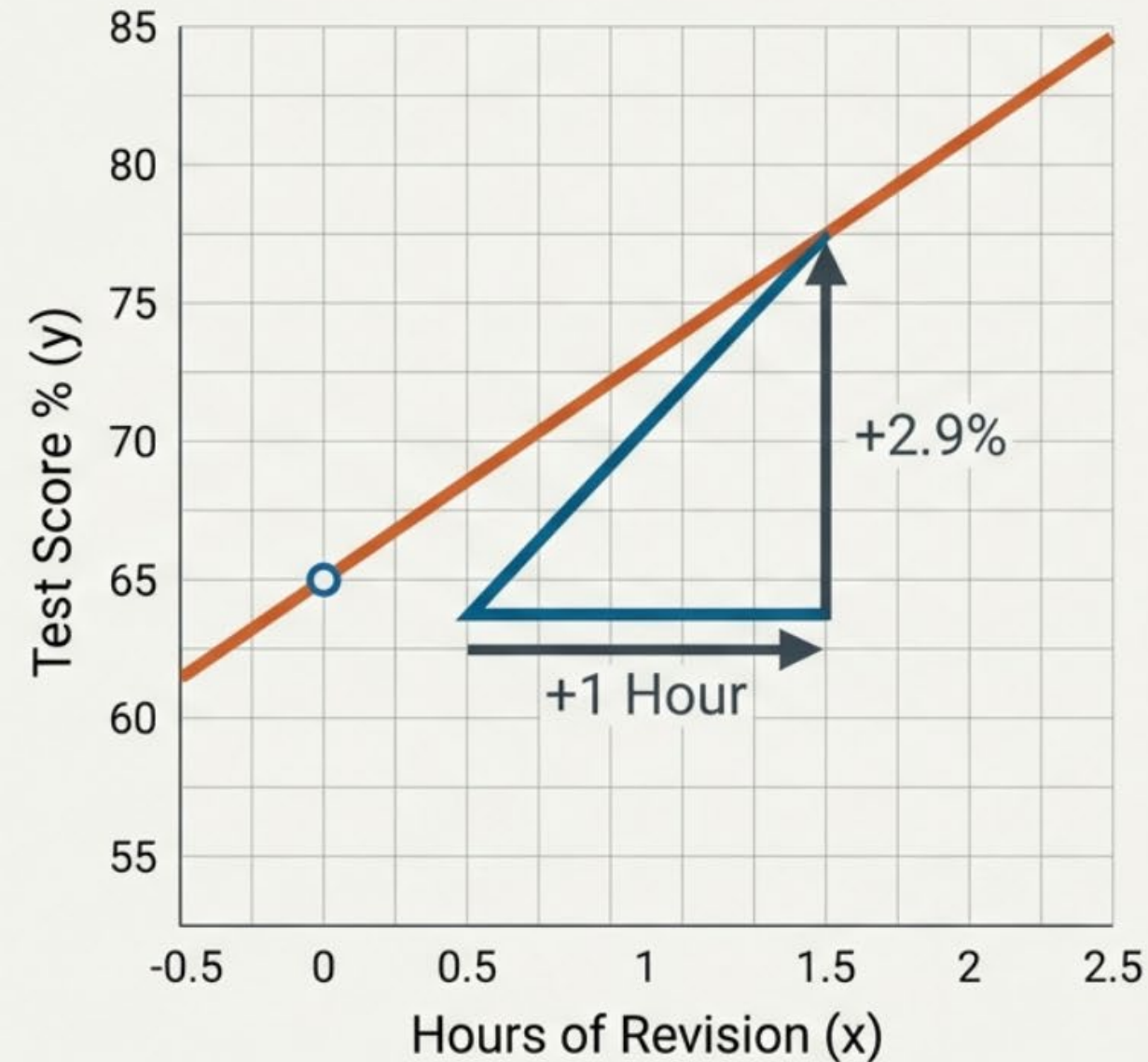


$$y = 65.5 + 2.9x$$

Interpretation: The value 65.5 represents the baseline.

Real World Meaning: A student who did 0 hours of revision would, on average, score 65.5% on the test.

Interpreting 'b': The Rate of Change



$$y = 65.5 + 2.9x$$

Gradient (b): 2.9

Interpretation: For every 1 unit increase in X, Y changes by 2.9.

Real World Meaning: Each extra hour of revision increases the test score by 2.9% on average.

$$\text{Gradient Units} = \frac{\text{Units of Y}}{\text{Units of X}} = \text{Percent per Hour}$$

Using the Model: Prediction

Scenario: A student revises for 7 hours. Predict the score.

Step 1: Identify $x = 7$

Step 2: Substitute.

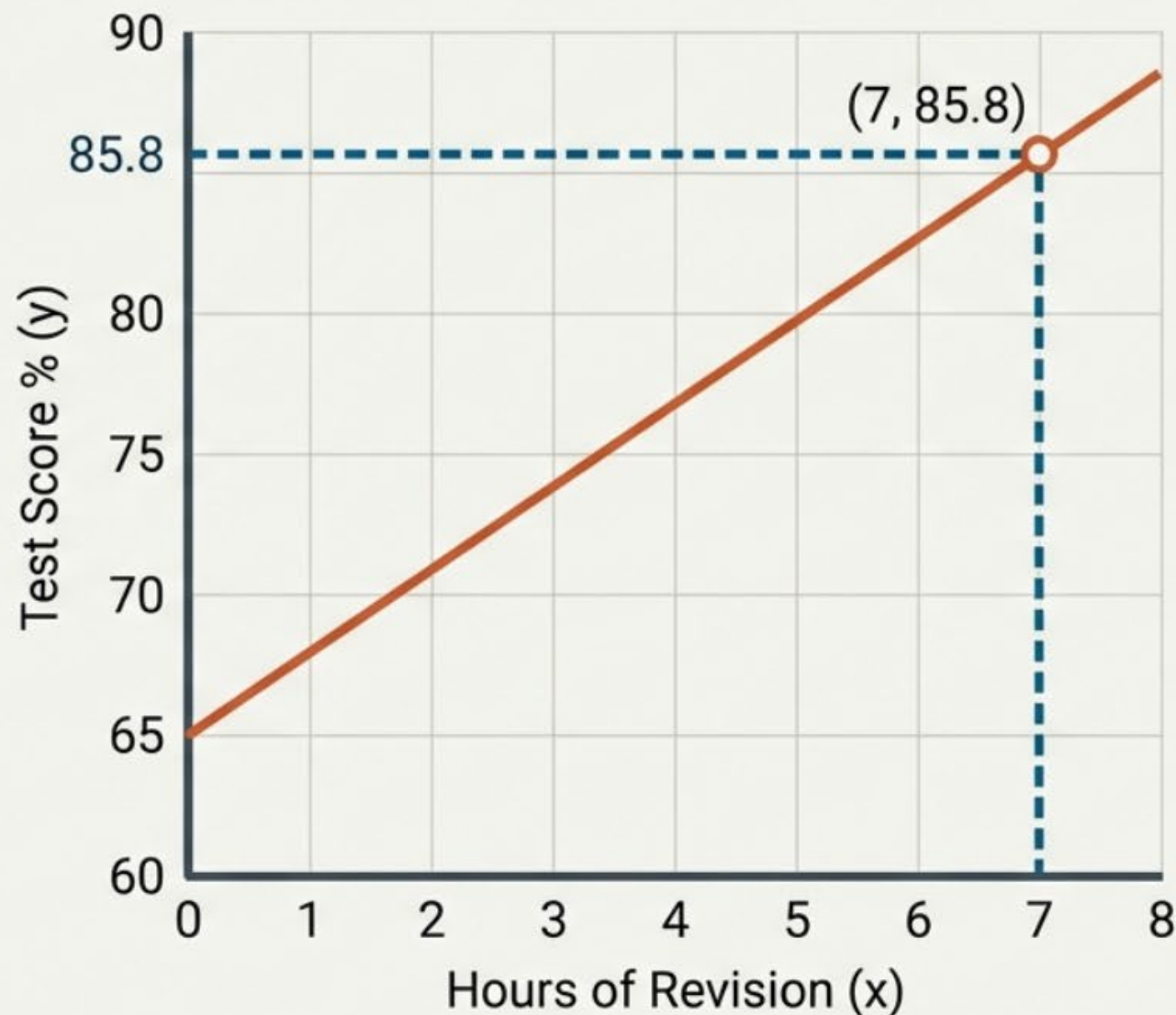
$$y = 65.5 + 2.9(7)$$

Step 3: Solve.

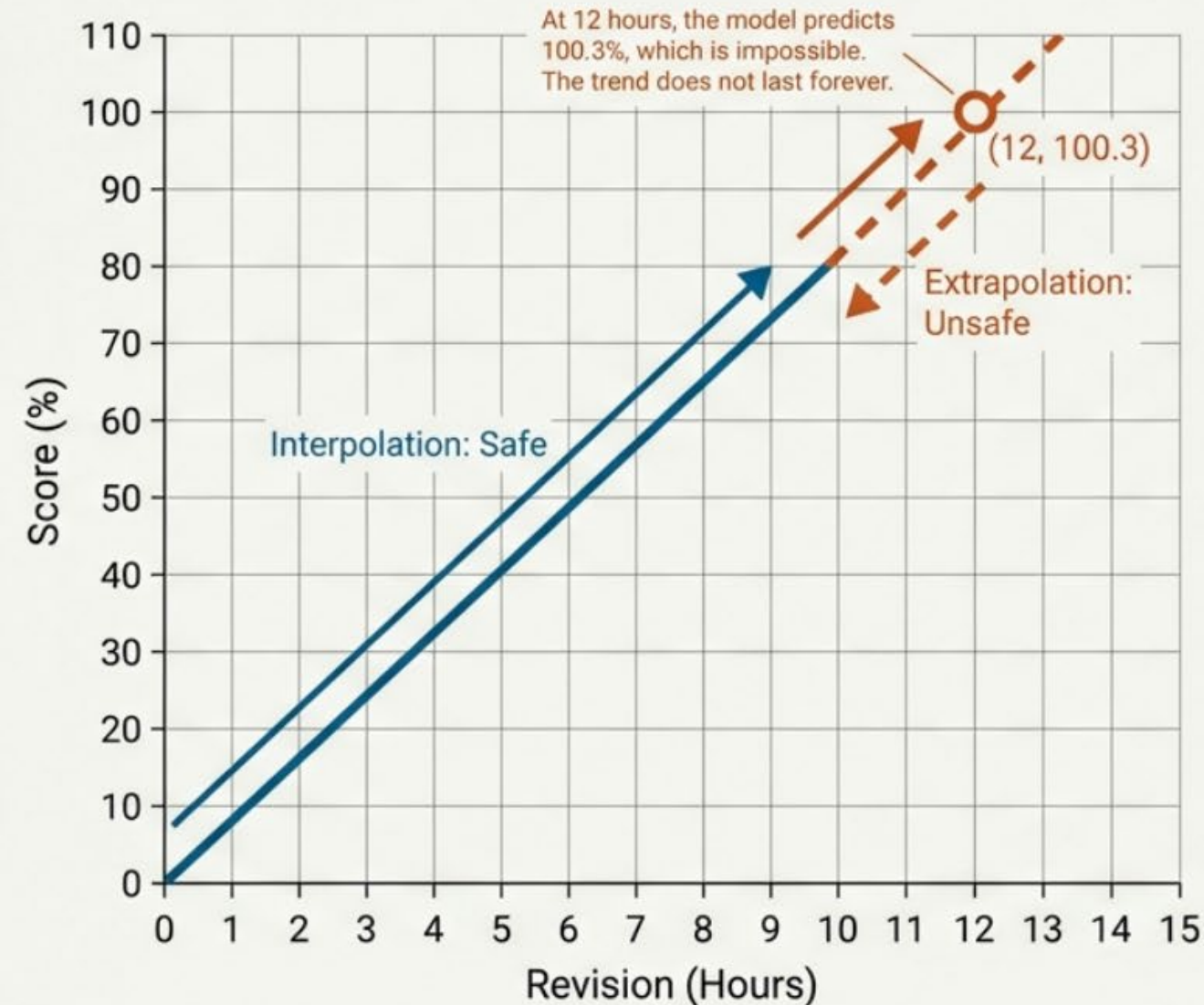
$$y = 65.5 + 20.3$$

Result: Predicted Score = 85.8%

Rule: Always predict Y from X.
Do not use the line in reverse.



The Danger Zone: Interpolation vs. Extrapolation



Interpolation:
Predicting inside the data range. Reliable.

Extrapolation:
Predicting outside the range. Unreliable.