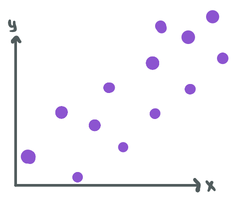
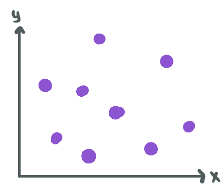
# Modeling Relationships of Bivariate Data

## Scatterplots Revistited

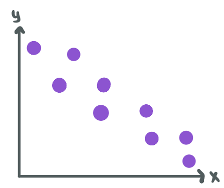
ex] State the type of Correlation from a scatter graph




Positive
Weak



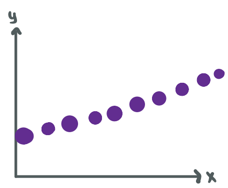

No
Correlation




Negative
Moderate

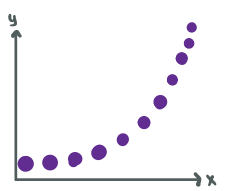
Sometimes it's difficult to determine the (+/-) Direction and Strength (strong, mod, weak) just by eye-balling it.

Linear

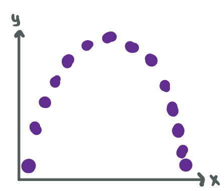



→ Constant Slope

Non-Linear




or



# Correlation vs. Causation

1. True or False?
   A. The IB grades achieved by a student are an effect of the predicted grades given by their instructors.
   B. The height of a person is a cause of their weight.
   C. The number of ice creams sold is a cause of high temperatures.
   D. Smoking is a cause of lung cancer.

A) False, correlation not causation

B) False, correlation not causation

C) False, correlation not causation

D) True

Qualitative relations

**Direction** - Positive, Negative

**Strength** - Weak, Moderate, Strong

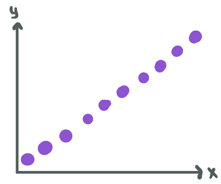
**Form** - Linear, Non-Linear

# Pearson's Product Moment Correlation Coefficient "r"

**r** – a measure of the linear correlation between two
two variables x and y. $-1 \le r \le 1$. r has no units

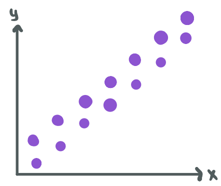
Correlation Coefficient is a measure of the strength of relationships

Positive Strong



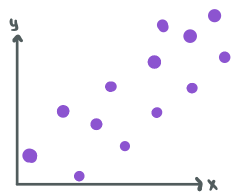

$r = 1$

Positive Moderate



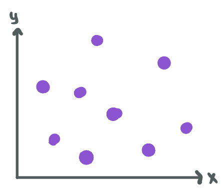

$r = 0.6$

Positive Weak



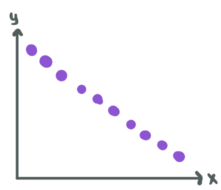

$r = 0.27$

No Correlation



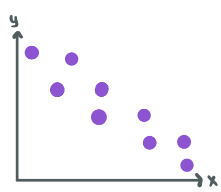

Non-Linear ⇒ Cannot be applied

Negative Strong




$r = -1$

Negative Moderate




$r = -0.52$

| $r$ | Correlation |
|---|---|
| $0 \le \lvert r \rvert < 0.25$ | Very weak |
| $0.25 \le \lvert r \rvert < 0.5$ | Weak |
| $0.5 \le \lvert r \rvert < 0.75$ | Moderate |
| $0.75 \le \lvert r \rvert < 1$ | Strong |

$$r = \frac{S_{xy}}{S_x S_y}$$, where

$S_{xy}$ - covariance (how x and y vary together)
$S_x$ - standard deviation of x
$S_y$ - standard deviation of y

$$S_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$S_y = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

You'll always be allowed your calculator

## Correlation Coefficient

Enter x-values in L1
Enter y-values in L2

$\boxed{\text{STAT}} \rightarrow \boxed{\text{CALC}} \rightarrow \boxed{8} \rightarrow \boxed{\text{Enter}} \rightarrow r = [\ ]$

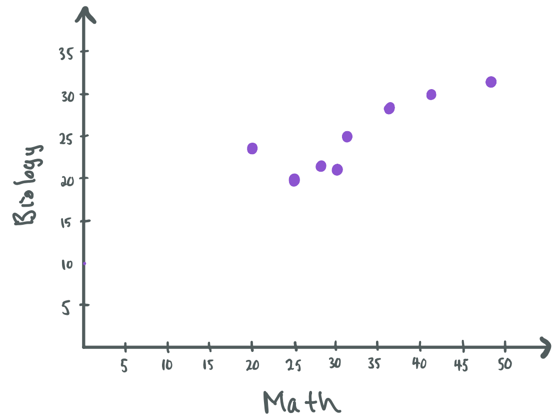
LinReg(a+bx)

? If Pearson's correlation coefficient ONLY models Linear Relationships
Is there another test to model the correlation of Non-Linear Relationships

ex) The table shows test results in math and biology for 8 students

| Math (x) | 20 | 25 | 28 | 30 | 32 | 42 | 48 |
|---|---|---|---|---|---|---|---|
| Biology (y) | 24 | 20 | 22 | 21 | 28 | 30 | 32 |

ⓐ Create a scatterplot




ⓑ What correlation type?

Positive Strong

ⓒ Find r

r= 0.83

ex) The table shows the height of plants

| Height (m) | 0 | 100 | 200 | 450 | 500 | 700 | 900 | 1000 |
|---|---|---|---|---|---|---|---|---|
| # of Plants | 1 | 2 | 5 | 8 | 8 | 10 | 12 | 13 |

ⓐ Find r

r= 0.98

ⓑ What correlation type?

Positive Strong

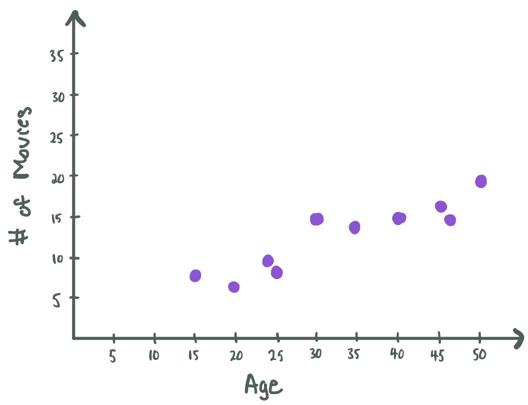
# Line of Best Fit $\hat{y}$

- a line drawn on a scatterplot to show the trend of the data points.

**Ex]** The table shows the age of a person (x), and the number of movies this person watched the last year (y)

| AGE | 15 | 20 | 24 | 25 | 30 | 35 | 40 | 45 | 46 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|
| # of MOVIES | 8 | 7 | 10 | 9 | 15 | 14 | 15 | 17 | 15 | 20 |

ⓐ Draw a Scatter Diagram




ⓑ Draw the line of best fit

Step 1 : Find the mean of the x-values $(\bar{x})$

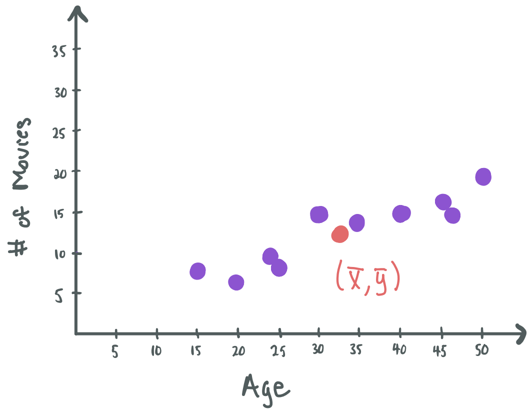
$$\bar{x} = \frac{15+20+24+25+30+35+40+45+46+50}{10} = \boxed{33}$$

Step 2 : Find the mean of the y-values $(\bar{y})$

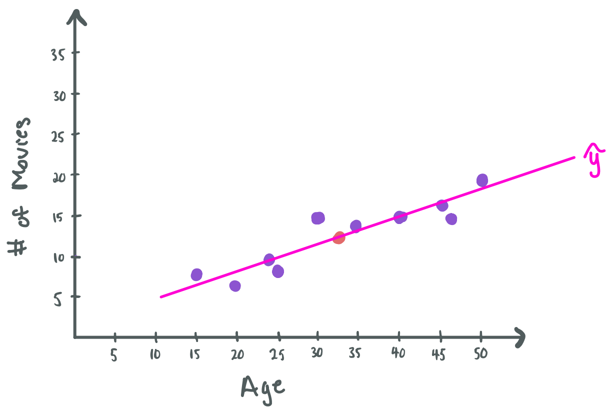
$$\bar{y} = \frac{8+7+10+9+15+14+15+17+15+20}{10} = \boxed{13}$$

Step 3 : Plot $(\bar{x}, \bar{y})$ on scatter graph = (33, 13)

☆ $(\bar{x}, \bar{y})$ is called the "Mean Point"




Step 4: Sketch line passing through the Mean Point such that an equal # of data points lie above and below the line




★ line of Best Fit is useful to make predictions

Ex] The table shows test results in math and biology for 8 students

| Math (x) | 20 | 25 | 28 | 30 | 32 | 37 | 42 | 48 |
|---|---|---|---|---|---|---|---|---|
| Biology (y) | 24 | 20 | 22 | 21 | 25 | 28 | 30 | 32 |

ⓐ Find the mean of the math test results

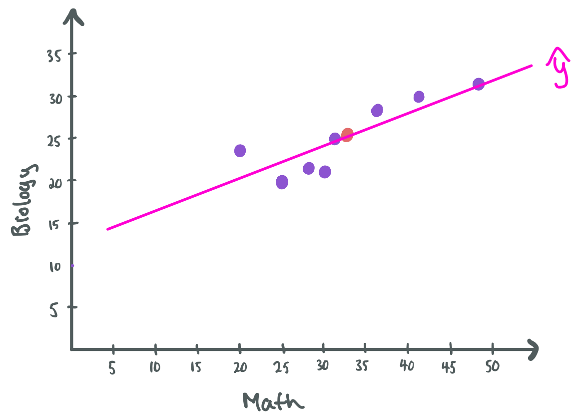
$\boxed{\bar{x} = 32.75}$

ⓑ Find the mean of the biology test results

$\boxed{\bar{y} = 25.25}$

ⓒ Plot + label the Mean Point and use it to draw the line of Best Fit by eye

(32.75, 25.25)

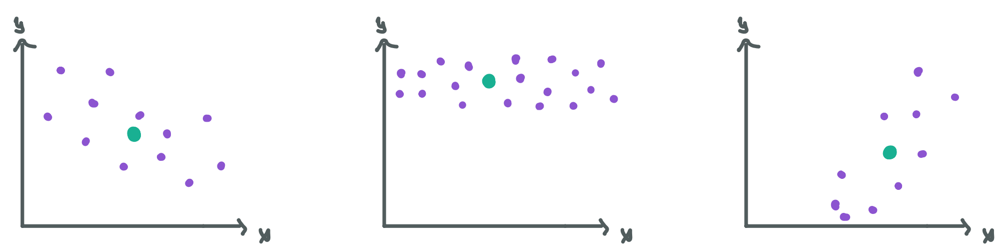



ⓓ Predict the biology test result for someone who got a 10 on their math test

$\boxed{17}$

Opener] Draw the line of best fit

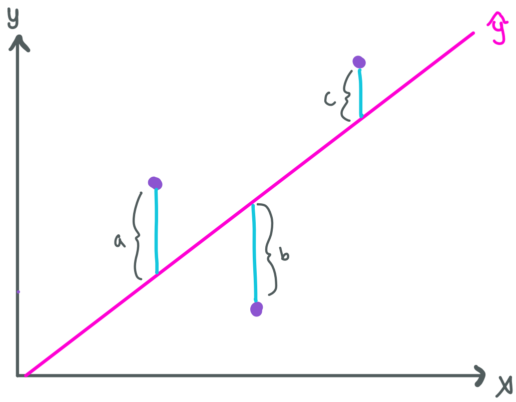
Mean Point



# Residual

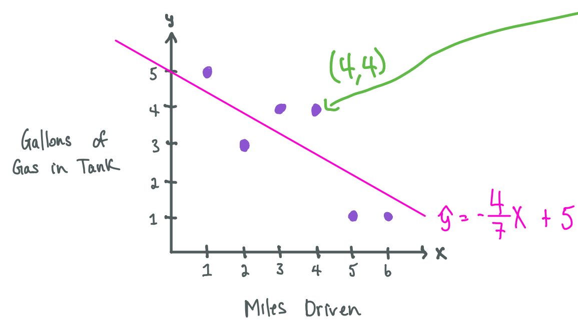
**Residual** - The difference btw the $\boxed{\text{actual}}$ y-value and the $\boxed{\text{predicted}}$ y-value. The error in our predictions when using the line of best fit

$$\text{Residual} = \text{actual} - \text{estimated} = y - \hat{y}$$




a, b, c are Residuals

**Residual Ex|** Find the Residual when 4 miles are driven




$\text{Residual} = y - \hat{y}$ → $\hat{y} = -\frac{4}{7}(4) + 5 = 2.71$

$= 4 - 2.71$

$\boxed{= 1.28}$

Find the Residual when 2 miles are driven

$$\text{Residual} = 3 - \left(-\frac{4}{7}(2) + 5\right)$$

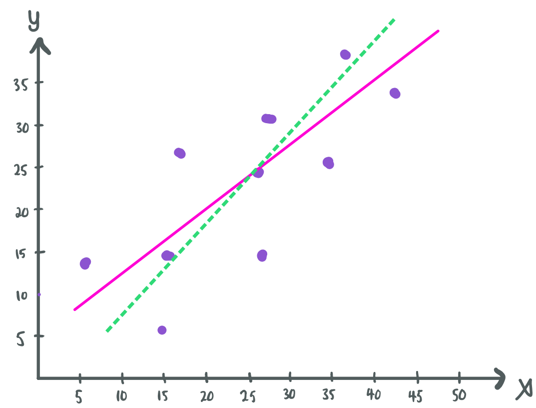
$$= \boxed{-0.85}$$

☆ Residuals can be negative

☆ To improve the accuracy, we want to minimize the sum of all Residuals. This is what we do when finding the least Squares Regression line ↓

**Least Squares Regression line** - the optimally placed line of Best Fit

It is line that minimizes the sum of all the Residuals




Least Squares Regression line

line of Best Fit

Ex] The table shows the height of plants

| Height (m) | 0 | 100 | 200 | 450 | 500 | 700 | 900 | 1000 |
|---|---|---|---|---|---|---|---|---|
| # of Plants | 1 | 2 | 5 | 8 | 8 | 10 | 12 | 13 |

ⓐ Graph the line of best fit on your calculator

ⓑ Find the equation for the line of best fit

# [Calc] Least Squares Regression Line

LinReg(a+bx) L1, L2, Y1 → ZOOM STAT

## Diagnostics

2nd → 0 → Diagnostic On → Enter → Enter

Catalog

## Stat Plot

2nd → Y= → Enter → Enter

STAT Plot

ON

## Graph

Enter x-values in L1
Enter y-values in L2

STAT → CALC → 8 → L1 → ,

LinReg(a+bx)

→ L2 → , → VARS → Y-VARS

→ Enter → Y1 → Enter

LinReg(a+bx) L1, L2, Y1 → Enter

y= a+bx
a=
b=
$r^2$=
r=

→ ZOOM → 9

ZOOM STAT

# Practice Problems

| | | | |
|---|---|---|---|
| Pg 270 | Exercise | 6B | Q 1,2,4 |
| Pg 272 | Exercise | 6C | Q 1 - 4 |
| Pg 274 | Exercise | 6D | Q 1-3 |
| Pg 276 | Exercise | 6E | Q 1-3 |
| 280 | E 6F | | Q 1-4 |