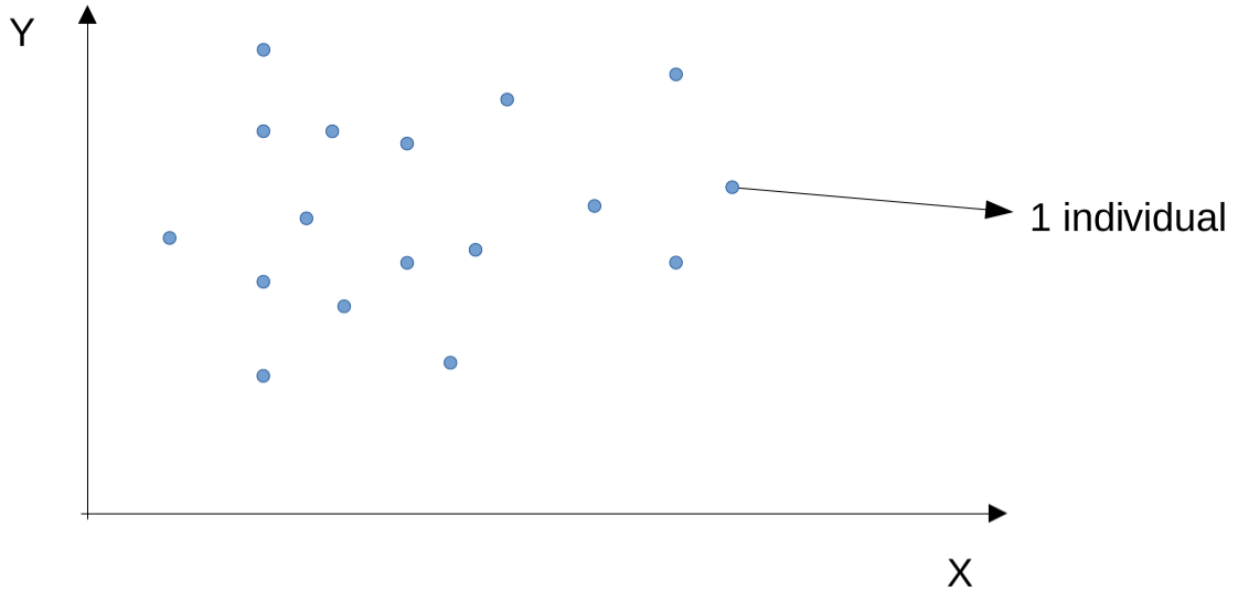
Study the link between two quantitative variables.

Examples :

- Is there a link between slepping duration and school performances ?
- Is there a link between newborns height and weight ?
- Is there a link between water supply and plant growth ?
- ...

# Graphical representation

**Linear relationship** between two quantitative variables measured on the **same individuals**!




Two random variables (X,Y) measured on a sample of n=17 individuals

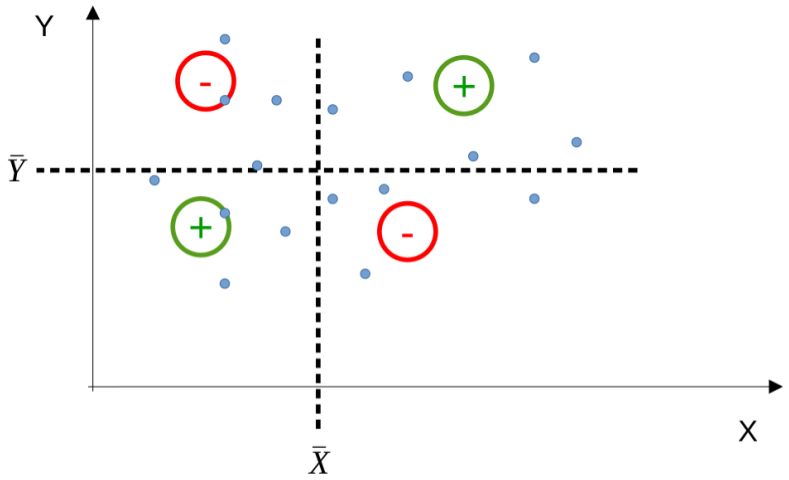
# Covariance

Variance

$$S_X^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Covariance

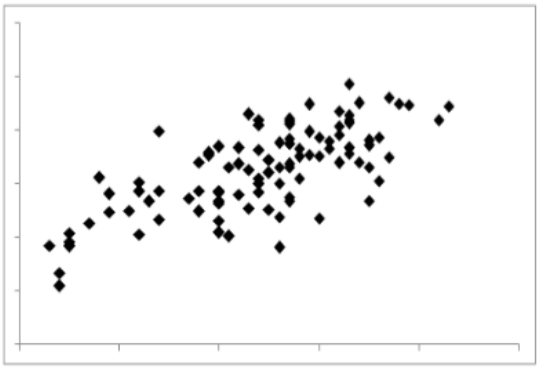
$$S_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$



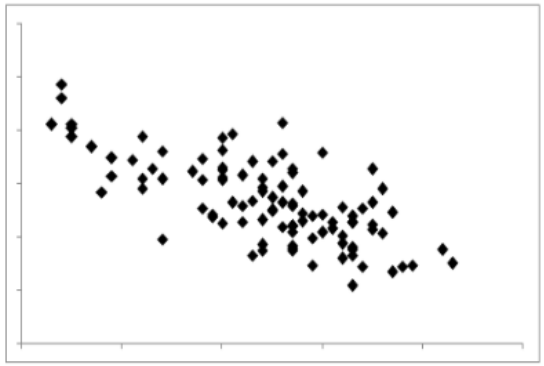

**Variance** : *V*ariance is a measure used to characterize the dispersion of a sample or a distribution.

**Covariance** : *C*ovariance measures the deviation from (statistical) independence between two random variables.

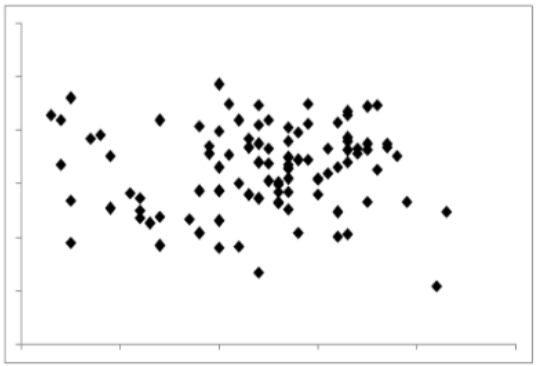
# Covariance



$S_{XY} > 0$



$S_{XY} < 0$



$S_{XY} = 0$

# Covariance

The correlation coefficient is a measure of covariance that is independent of the units in which X and Y are expressed.

Variance

$$S_X^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

unité de $X^2$

Covariance

$$S_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Unité de X.Y

## Coefficient de corrélation de Pearson

Population

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X . \sigma_Y}$$

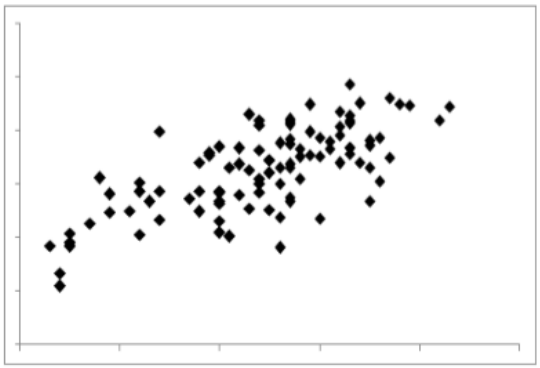
ρ = rhô
σ = sigma

Sample

$$r_{XY} = \frac{S_{XY}}{S_X . S_Y}$$

Sans unité

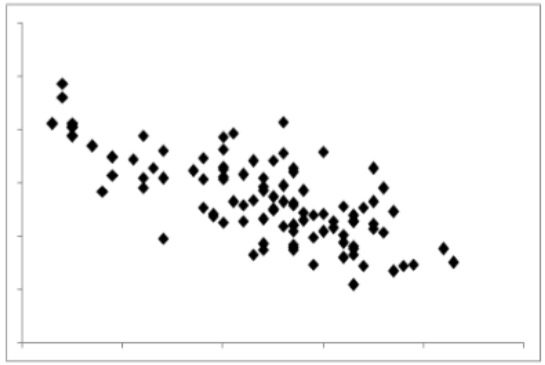
# Pearson's Correlation Coefficient



$\sigma_{XY} > 0$

$0 < \rho_{XY} < 1$

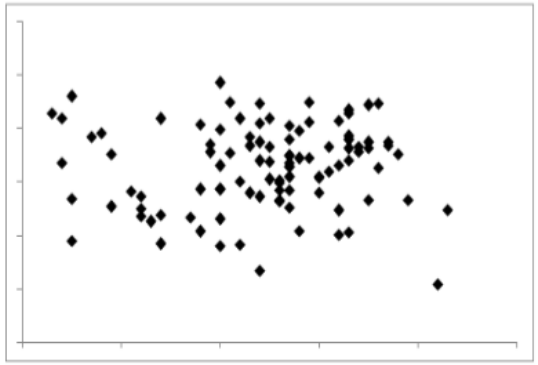
Relation linéaire positive



$\sigma_{XY} < 0$

$-1 < \rho_{XY} < 0$

Relation linéaire négative



$\sigma_{XY} = 0$

$\rho_{XY} = 0$

Pas de relation
Indépendence

# Correlation Test

Biological question: Is there a relationship between X and Y?

**1. The model**

X and Y are a Gaussian pair of variables:

- X N($\mu_X$ ; $\sigma^2_X$) for all Y
- Y N($\mu_X$ ; $\sigma^2_X$) for all X

**2. The hypotheses**

- H0: X and Y are independent, $\rho_{XY} = 0$ (but $r_{XY}$ may differ from 0)
- H1: X and Y are not independent: $\rho_{XY} \neq 0$ (two-sided), $\rho_{XY} > 0$ (one-sided), $\rho_{XY} < 0$ (one-sided)

# Correlation Test

Biological question: Is there a relationship between X and Y?

3. **Test statistic and its distribution under H0**

$$T = \frac{R_{XY} . \sqrt{n-2}}{\sqrt{1 - R_{XY}^2}} \sim_{H0} T_{(n-2)}$$

4. **Rejection region and threshold calculation**

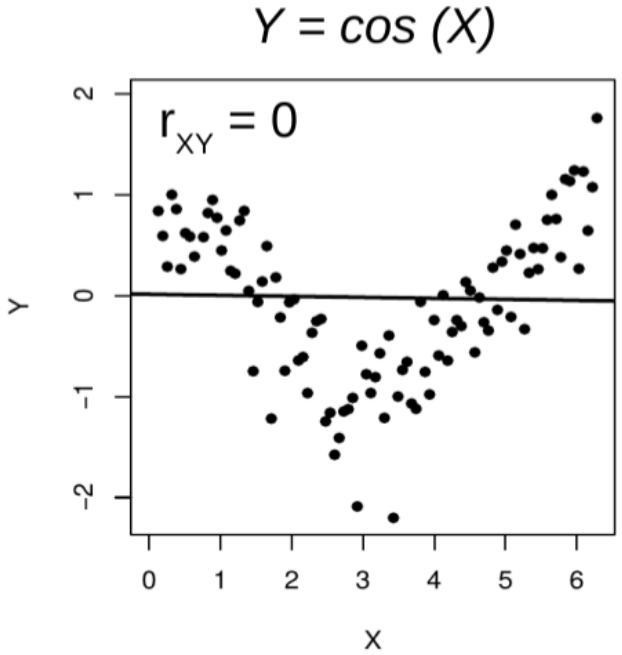
5. **Choose risk $\alpha = 0.05$ by default**

6. **Calculation of $T_{obs}$**

7. **Calculation of the p-value and conclusion**

# Correlation Test

Two variables can be linked by a non-linear relationship.




Always graphically represent the data beforehand.
If the relationship does not seem linear:

- Perform a data transformation (e.g. log)
- Use Spearman's correlation coefficient

# Spearman's Correlation Test

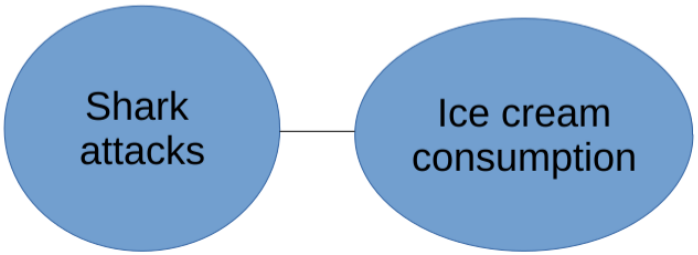
In cases where:

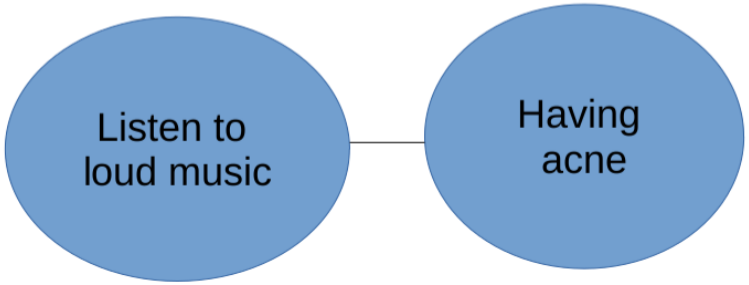
- X and Y are not a Gaussian pair
- The samples are small ($<30$)
- A non-linear relationship is expected

For each variable, the ranks of individuals are calculated, and two new variables are considered: $R_X$ and $R_Y$

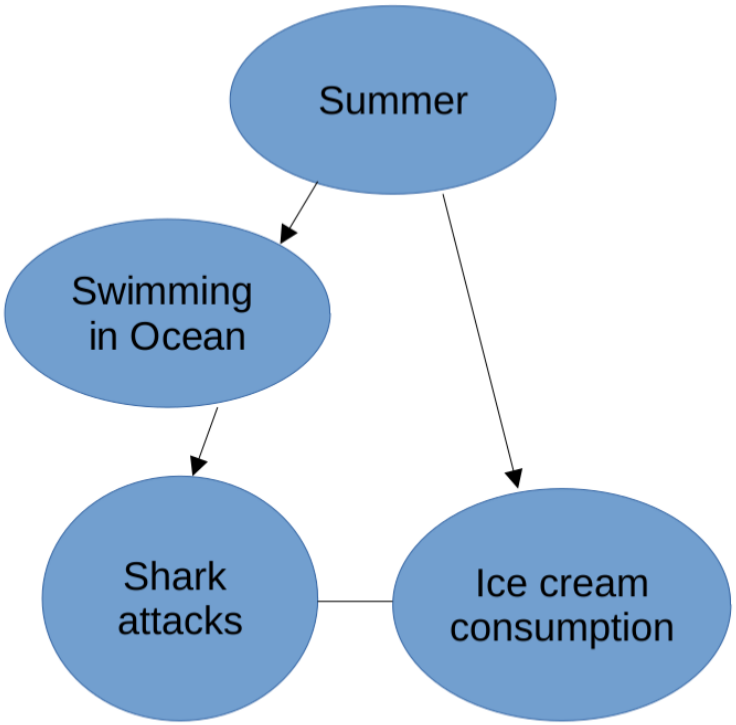
The Pearson correlation test is then performed on $R_X$ and $R_Y$

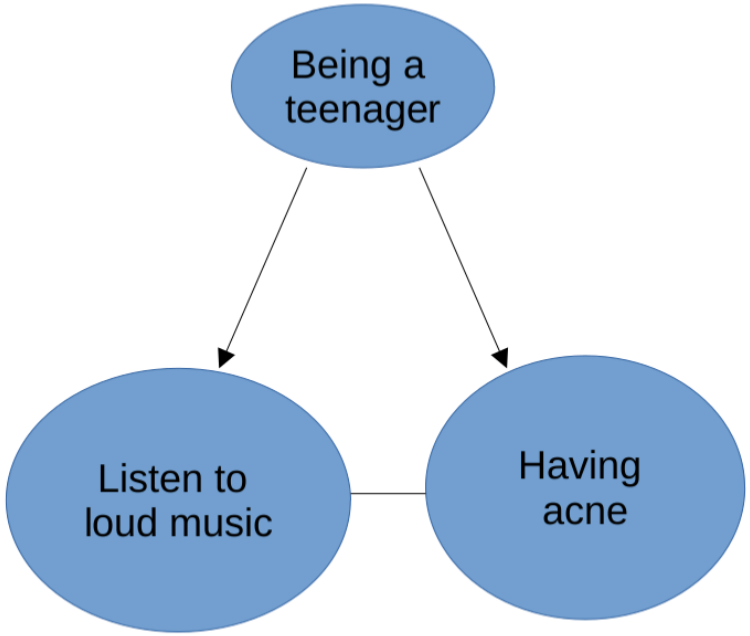
# Correlation $\neq$ Causality



# Correlation $\neq$ Causality




A causal relationship is always difficult to establish and prove.

# Exercise - Part 1: Correlation

The dataset *yields* provides the yield of a wheat variety grown in different environments where soil and climate variables were collected.

- Load the dataset and identify the number of variables and individuals
- Visualize the distribution of each variable and assess their normality (*hist*)
- Create a scatterplot matrix to visualize the relationships between all pairs of variables (*plot*)
- Calculate the Pearson correlation coefficient for each pair of variables (*cor*)
- Which pairs of variables seem dependent? Independent?
- Now, perform statistical tests to identify significantly correlated variables (*cor.test*). Interpret the test output.

# Linear Regression

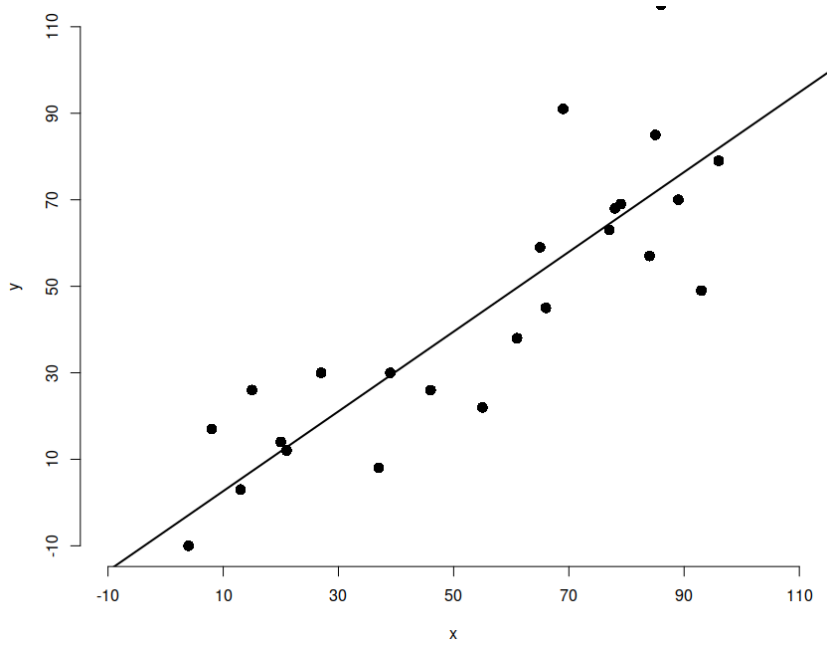
We want to predict the values of a variable Y (**response variable**) based on the observation of another variable X (**explanatory variable**).

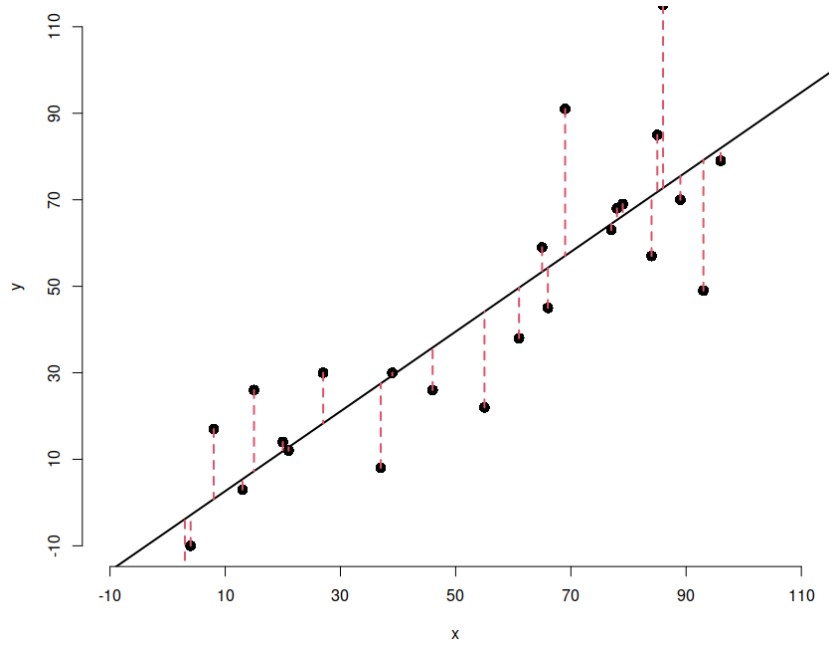
The model: $Y_i = a + bx_i + \epsilon_i$ where $\epsilon \sim N(0, \sigma^2)$

We aim to identify the parameters $a$ and $b$ to obtain the line that best fits the scatterplot.

$\epsilon$: epsilon; $\sigma$: sigma

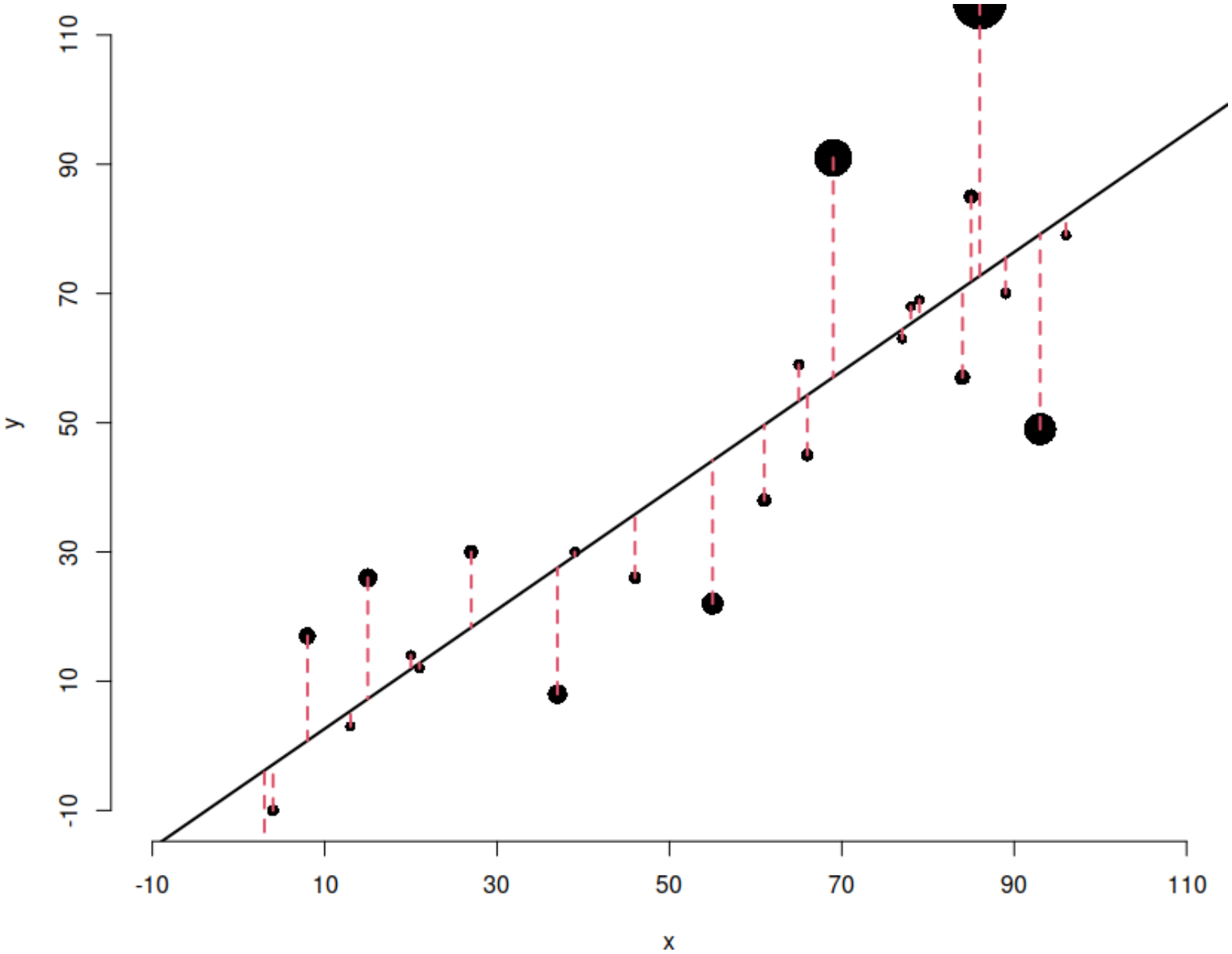
# Linear Regression




We aim to minimize the squared error: $\sum_{i=1}^{n}(Y_i - (a + bx_i)^2)$

# Linear Regression



# Linear Regression

If the model is correct, the residuals should represent random variations:

- Centered around zero
- Normally distributed (QQplot)
- Independent of predicted values (residuals vs fitted values plot)

# Linear Regression

*a* is the intercept of the Y-axis. It is the predicted value of y when x=0, which may or may not have biological significance.

*b* is the slope of the line.

For each of these coefficients, a test can be performed:

- H0: $a = 0$ / H1: $a \neq 0$
- H0: $b = 0$ / H1: $b \neq 0$

The coefficient of determination ($r^2$) represents the percentage of variance in Y explained by variations in X.

$0 < r^2 < 1$

$$\frac{Var(Model)}{Var(tot)} = \frac{S_{XY}^2}{S_X^2 S_Y^2} = r_{XY}^2$$

The closer $r^2$ is to 1, the better the model.

# Exercise - Part 2: Regression

In this part, we will try to identify a way to predict the yield of a plot based on environmental variables.

- Among the provided environmental variables, which would you use to predict yield?
- Apply the model to make this prediction and identify the equation to make predictions (*lm* and *summary*).
- What proportion of the yield variance is predicted by your model?
- Are there other variables you would use to improve your prediction? Try adding them and compare the $r^2$ of the different models.
- In conclusion, which variable(s) do you need to predict wheat yield, and what equation would you apply?