

A decorative border surrounds the central text, featuring various hand-drawn icons. On the left, there is a heart, a lightbulb, a star, a musical note, a sun, a leaf, and a large letter 'A'. On the right, there is a book, a ruler, a globe, a protractor, a heart, a large letter 'A', and a ruler. At the bottom, there is a book, a large number '3', a document, a large number '2', a heart, a pushpin, a beaker, a large star, and a large letter 'C'.

AGR5201

ADVANCED STATISTICAL METHODS

Regression and Correlation Analysis

Topic outline

1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

Correlation coefficient, r

4.0 Summary | Regression vs. correlation



1.0 Regression and correlation analysis



So far....

- We have been estimating differences caused by application of various treatments, and determining the probability that an observed difference was due to chance → ANOVA
- But we have not learned anything about how two (or more) variables are related

1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

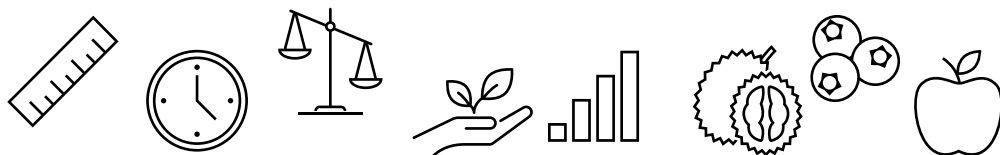
Correlation coefficient, r

4.0 Summary | Regression vs. correlation

1.0 Regression and correlation analysis

Types of variables in biology

- **Treatments** → fertilizer rates, varieties, and weed control methods which are the primary focus of the experiment
- **Environmental factors** → rainfall and solar radiation which are not within the researcher's control
- **Responses** → represent the biological and physical features of the experimental units that are expected to be affected by the treatments being tested.
 - E.g: yield or weight (affected by treatment)



1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

Correlation coefficient, r

4.0 Summary | Regression vs. correlation

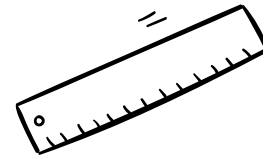
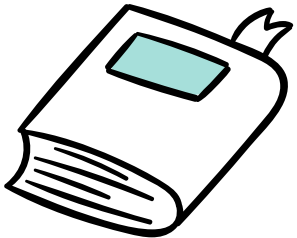
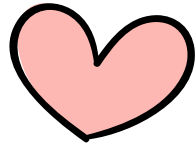
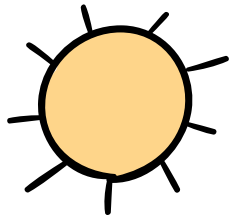
1.0 Regression and correlation analysis

Common association within ANOVA:

- Agronomic experiments frequently consist of different levels of one or more **quantitative variables**:
 - Varying amounts of fertilizer
 - Several different row spacings
- It would be useful to develop an equation to describe the relationship between plant response and treatment level
 - the response could then be specified for not only the treatment levels that was tested but for all other intermediate points within the range of those treatments
- The simplest form of response is a straight line

1.0 Regression and correlation analysis
Introduction
2.0 Regression analysis
Regression model, equation and plot
Coefficient of determination, r^2
Simple linear regression
Example
3.0 Correlation analysis
Relationship between variables
Correlation coefficient, r
4.0 Summary Regression vs. correlation

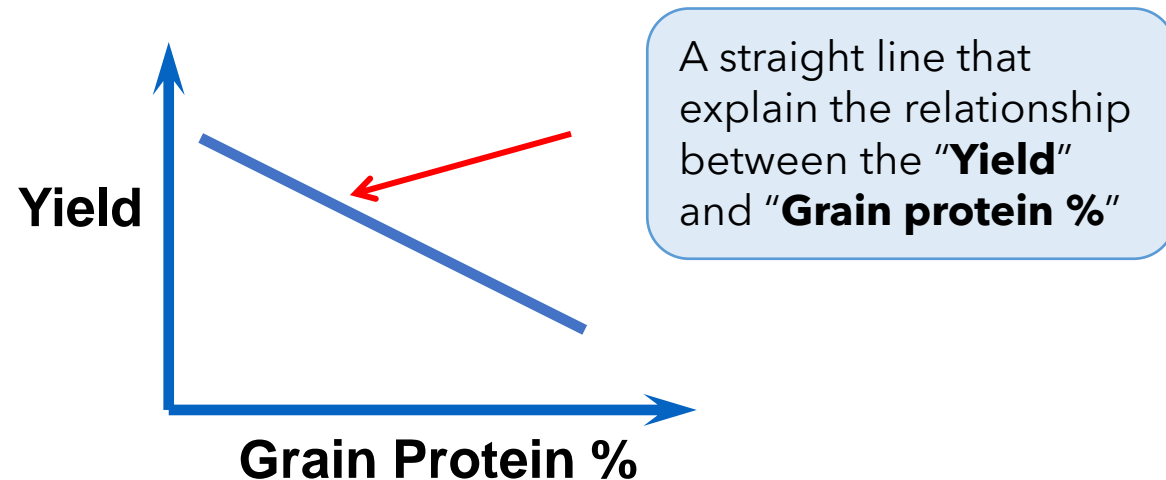
REGRESSION ANALYSIS



2.0 Regression analysis

Regression

- The modeling of the relationship between a response variable and a set of explanatory variables.
- Allows:
 - the user to determine which of the explanatory variables have an effect on the response.
 - the users to explore what happens to the response variable for specified changes in the explanatory variables.



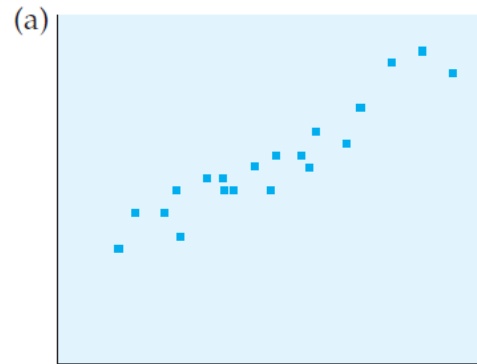
Explain how one variable is related to another
As you change one variable, how are others affected?

- ✓ May want to
 - Develop and test a model for a biological system
 - Predict the values of one variable from another

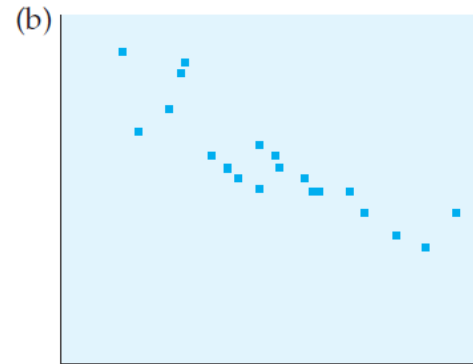
2.0 Regression analysis

Types of relationship

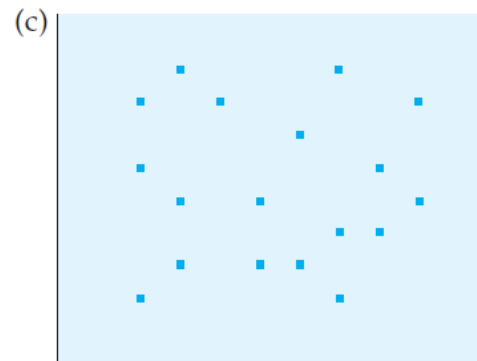
Positive linear relationship



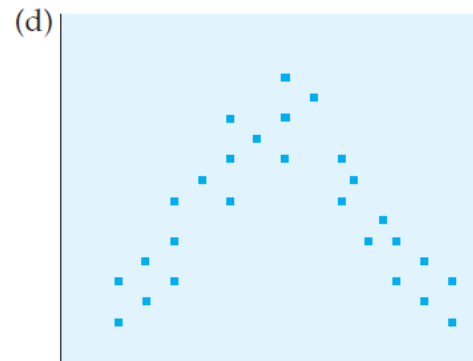
Negative linear relationship



No relationship



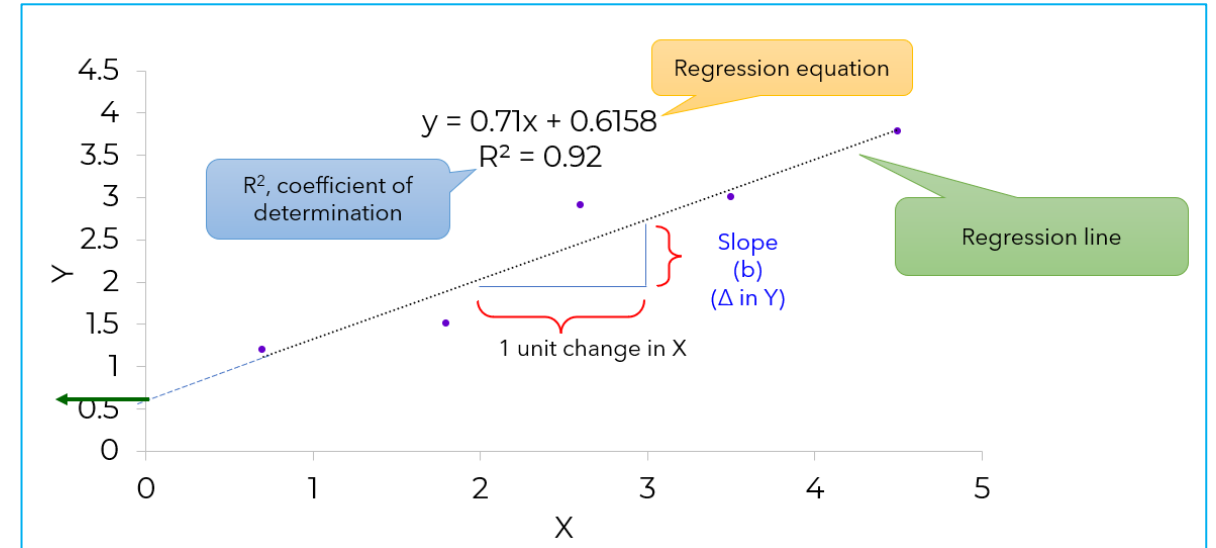
Nonlinear relationship (curvilinear)



2.0 Regression analysis

Components in regression analysis

- i. Two quantitative variables. Examples:
 - Weight and height
 - Age and blood pressure
 - Dry weight and plant height
- ii. Regression equation
- iii. Regression plot (line and equation)
- iv. Coefficient of determination, R^2



1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

Correlation coefficient, r

4.0 Summary | Regression vs. correlation

2.0 Regression analysis

Response and explanatory variables

- **Response variable:**
 - The outcome variable on which comparisons are made
 - It is a dependent variable (it depends on the other variables)
 - Quantitative variable
- **Explanatory variables:**
 - It is an independent variable (does not depend on the other variable)
 - Also known as predictor variable
 - Qualitative (categorical): the groups to be compared
 - Quantitative: the change in numerical values to be compared
- Regression analysis examines how the **response variable** depends on or explained by the **explanatory variables**

1.0 Regression and correlation analysis
Introduction
2.0 Regression analysis
Regression model, equation and plot
Coefficient of determination, r^2
Simple linear regression
Example
3.0 Correlation analysis
Relationship between variables
Correlation coefficient, r
4.0 Summary Regression vs. correlation

2.0 Regression analysis

Summary of regression analysis

Types of regression	Conditions
Univariate	Only one quantitative response variable
Multivariate	Multiple quantitative response variables
Simple	Only one predictor variable
Multiple	Multiple predictor variables
Linear	All parameters enter the equation linearly
Nonlinear	The relationship between predictor and response is nonlinear
Analysis of variance	All predictors are qualitative (category) variables
Analysis of covariance	Some predictors are qualitative and some are quantitative
Logistics	The response variable is qualitative (category i.e. yes or no)

2.0 Regression analysis

Linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where:

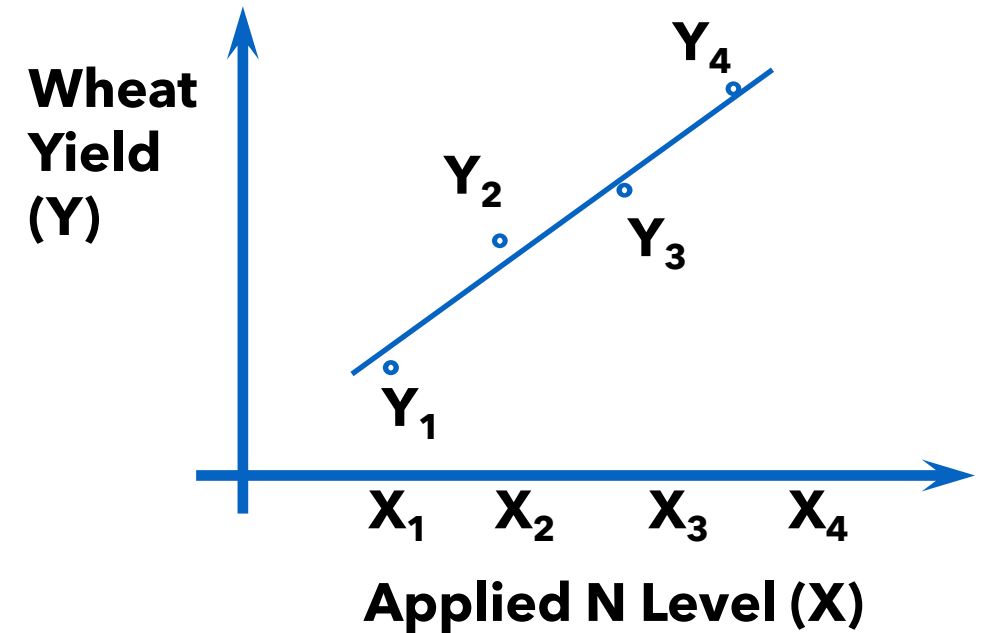
Y = wheat yield (dependent variable)

X = nitrogen level (independent variable)

β_0 = intercept (value of Y when X=0)

β_1 = Slope (change in Y for every unit change in X)

ε = random error



- ✓ Regression → Choose a line that minimizes deviation of observed values from the line (predicted values)

2.0 Regression analysis

Types of regression models

Model I

- Values of the independent variable X are controlled by the experimenter
- Assumed to be measured without error
- We measure response of the independent variable Y to changes in X
- Example: Treatment (quantitative) vs. response (rate of nitrogen vs. yield)

Model II

- Both the X and the Y variables are measured and subject to error (e.g., in an observational study)
- Either variable could be considered as the independent variable; choice depends on the context of the experiment
- Often interested in correlations between variables
- May be descriptive, but might not be reliable for prediction
- Example: between two measured variables (plant height vs. plant dry weight)

2.0 Regression analysis | Types of regression

Simple linear regression

- Has only one independent variable (predictor)
- The simplest relationship between two variables is a straight line
- "Simple" indicates that there is only one independent variable.
- "Linear" indicates the nature of the model type.

$$Y = a + bX$$

Yield intercept $b = \text{slope}$ fertilizer

Multiple linear regression

- An extension of simple regression that has more than one independent variable (predictors)

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

Yield intercept $b_1 - b_n = \text{slope coefficients}$ fertilizer irrigation other factors

1.0 Regression and correlation analysis
Introduction
2.0 Regression analysis
Regression model, equation and plot
Coefficient of determination, r^2
Simple linear regression
Example
3.0 Correlation analysis
Relationship between variables
Correlation coefficient, r
4.0 Summary Regression vs. correlation

2.0 Regression analysis

Regression equation

- The equation of the straight line through the data is known as a regression equation.
- Regression equation is formulated using a least square (LS) method
- LS method is a procedure that minimizes the vertical deviations of plotted points surrounding a straight line

Linear regression equation formula:

$$\hat{Y} = a + bX$$

Where,

a = Y-intercept (value of Y when $X = 0$)

b = slope (a change in Y for a change in 1 unit of X)

X = Variable X

\hat{Y} = Estimated value of Y

Regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

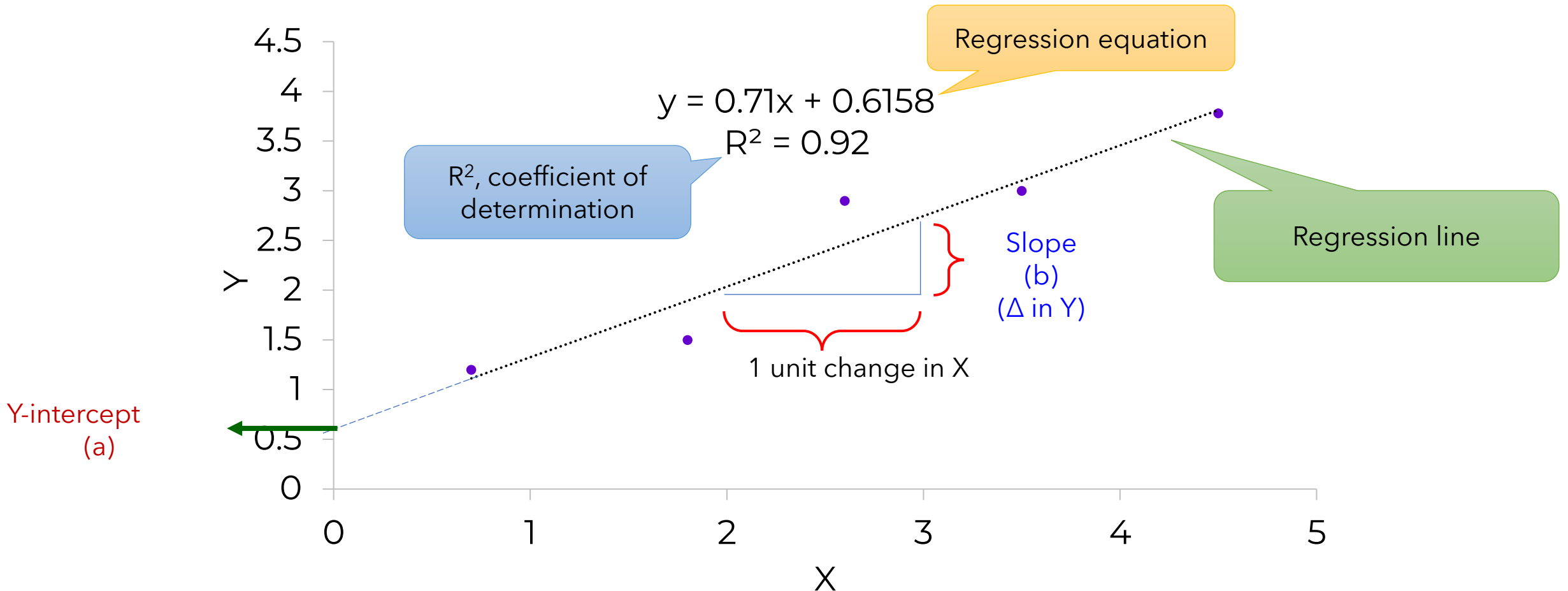
Relationship between variables

Correlation coefficient, r

4.0 Summary | Regression vs. correlation

2.0 Regression analysis

Regression plot (line, equation and coefficients)



2.0 Regression analysis

Regression line

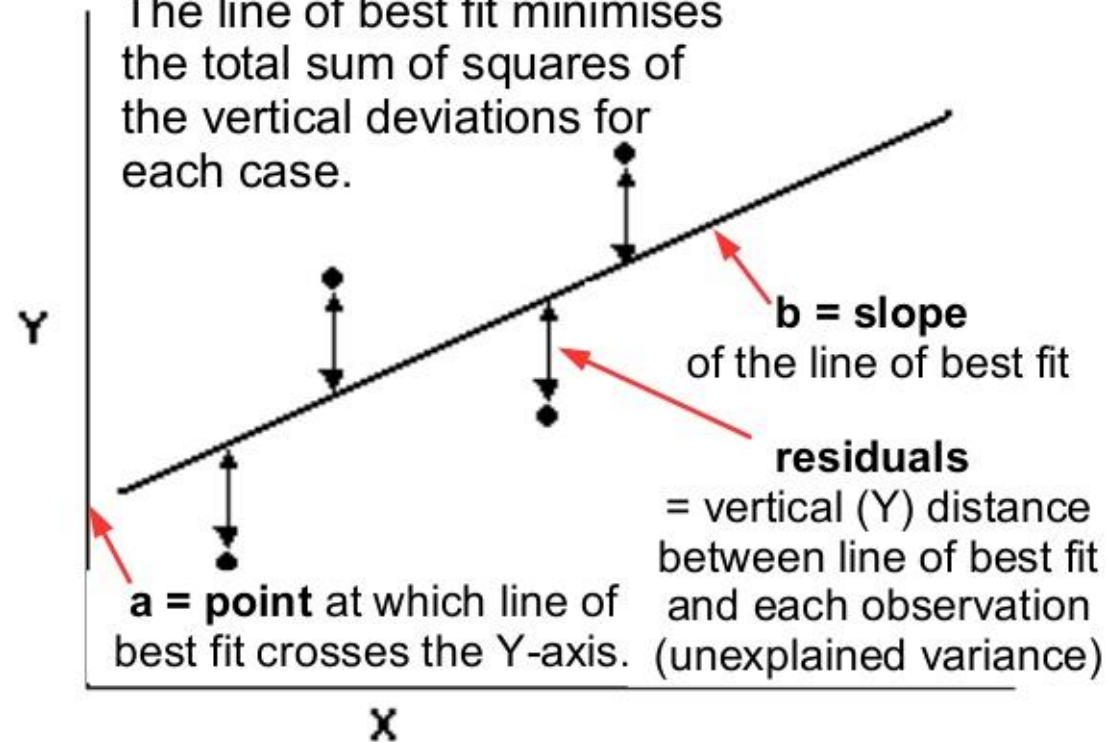
- Calculates the “best-fit” line for a certain set of data.
- The regression line makes the sum of the squares of the residuals smaller than for any other line.
- We attempt to minimize the residuals (error term in the regression model)
- $\text{Residuals} = \text{Observation} - \text{grand mean}$

2.0 Regression analysis

Least square criterion

Least squares criterion

The line of best fit minimises the total sum of squares of the vertical deviations for each case.



2.0 Regression analysis

Regression equation | Coefficient interpretation

$$\hat{y} = 0.71(x) + 0.6158$$

- The numbers in the equation convey some information about the relationship between variables.
- The interpretation of intercept and slope values:
 - **a (intercept)** → y value when x = 0, is 0.6158
 - **b (slope)** → y changes 0.71 unit when x changes by 1 unit
 - (in this case, the relationship is positive, thus it means the y increases 0.71 unit cm for 1 unit increase in x)



Linear regression equation:

$$\hat{Y} = bX + a$$

1.0 Regression and correlation analysis
Introduction
2.0 Regression analysis
Regression model, equation and plot
Coefficient of determination, r^2
Simple linear regression
Example
3.0 Correlation analysis
Relationship between variables
Correlation coefficient, r
4.0 Summary Regression vs. correlation

2.0 Regression analysis

Regression equation | Prediction

- Regression analysis allows a prediction of one variable knowing another variable
- For example: By knowing x , we can predict the value of y .
- Regression analysis deals with:
 - Regression equation - a mathematical equation that indicates the relationship between variables
 - Regression line - a line through a cloud of datapoints on scatterplot

1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

Correlation coefficient, r

4.0 Summary | Regression vs. correlation

2.0 Regression analysis | Simple linear regression

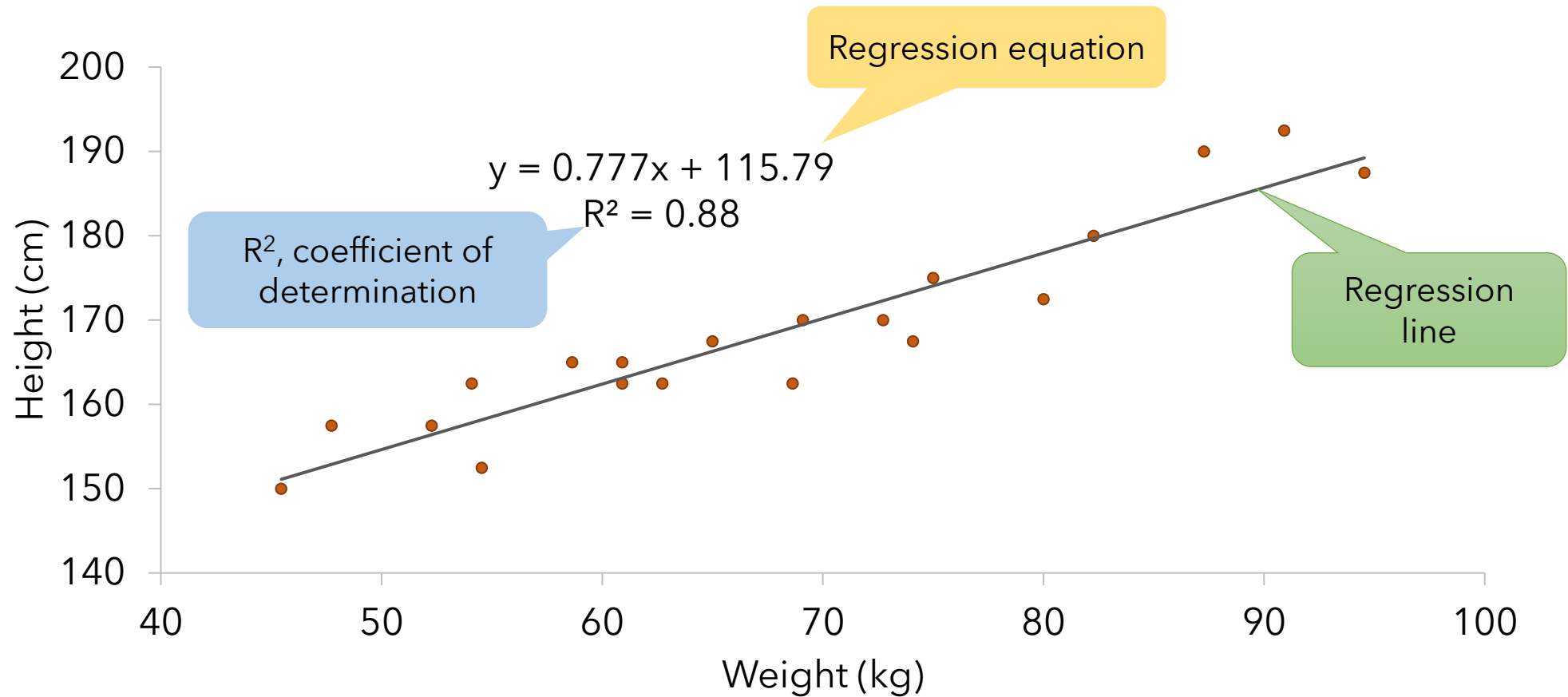
Example: Height vs. weight

- A researcher would like to study the relationship between weight and height and make a prediction of height based on weight
- The table on the right shows data on height and weight of 20 individuals
- Conduct a regression analysis to determine the type of relationship and develop a regression equation for prediction

Individual	weight	height
1	45.45	150
2	54.55	152.5
3	47.73	157.5
4	52.27	157.5
5	54.09	162.5
6	60.91	162.5
7	58.64	165
8	65	167.5
9	68.64	162.5
10	74.09	167.5
11	72.73	170
12	80	172.5
13	75	175
14	82.27	180
15	87.27	190
16	94.55	187.5
17	90.91	192.5
18	69.09	170
19	60.91	165
20	62.73	162.5

2.0 Regression analysis

Example: Regression line, equation and r^2



2.0 Regression analysis

Example: Interpretation of the coefficients

$$\widehat{Height} = 0.776 (weight) + 115.86$$

- The numbers in the equation convey some information about the relationship between variables.
- The biological interpretation of intercept and slope values:
 - **a (intercept)** : the height when weight = 0 is 115.86 cm (might not make sense, but that is how we interpret it)
 - **b (slope)**: the height changes 0.776 cm for every change of 1 kg in weight
 - (in this case, the relationship is +ve, thus it means the height increases 0.776 cm for 1 kg increase in weight.)



Linear regression
equation: $\hat{Y} = bX + a$

2.0 Regression analysis

Example: Prediction

- Regression analysis allows a prediction of one variable knowing another variable
- For example: By knowing one's height, we can predict the weight of that person.
- Regression analysis deals with:
 - Regression equation – a mathematical equation that indicates the relationship between variables
 - Regression line – a line through a cloud of datapoints on scatterplot

1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

Correlation coefficient, r

4.0 Summary | Regression vs. correlation

2.0 Regression analysis

Regression equation | Predicting Y based on X value

$$\hat{Y} = 0.776 X + 115.86$$
$$\widehat{Height} = 0.776 (weight) + 115.86$$

- Using the regression equation above, we can predict (estimate) the value of Y using any value X.
- **Please note that the value of X must be between the range of X values of the data collected. It is not appropriate to predict Y from values outside the range of X value of the dataset used to construct the regression equation.
- To estimate the height of individual with weight of 56 kg:

$$\widehat{Height} = 0.776 (56) + 115.86 = 43.456 + 115.86$$
$$\underline{\widehat{Height} = 159.32 \text{ cm}}$$

1.0 Regression and correlation analysis
Introduction
2.0 Regression analysis
Regression model, equation and plot
Coefficient of determination, r^2
Simple linear regression
Example
3.0 Correlation analysis
Relationship between variables
Correlation coefficient, r
4.0 Summary Regression vs. correlation

2.0 Regression analysis

Coefficient of determination, R^2

- R^2 is a value indicate how much of the variation in Y is explained by the regression line, or how good is the model explain your data.

- **Formula of R^2 :**

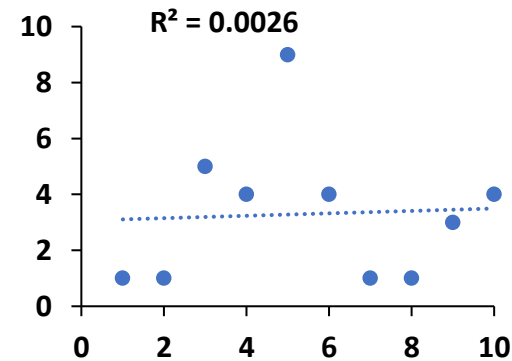
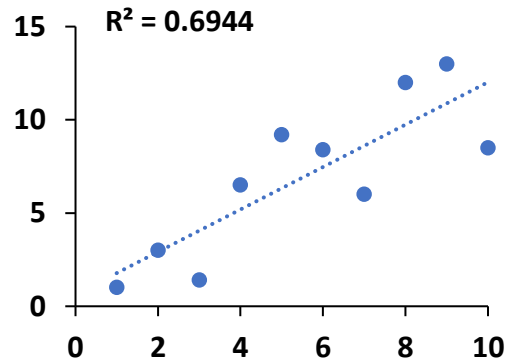
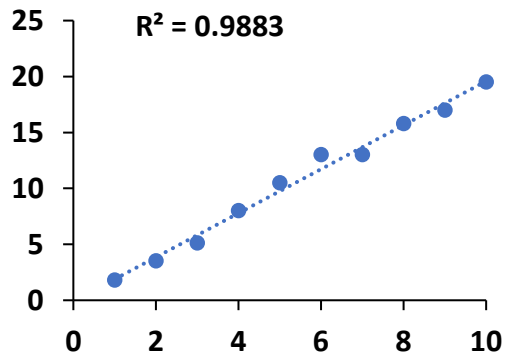
$$R^2 = 1 - \frac{\text{SSE}}{\sum (y_i - \bar{y})^2} = \frac{\sum (y_i - \bar{y})^2 - \text{SSE}}{\sum (y_i - \bar{y})^2} = \frac{\text{SSR}}{\sum (y_i - \bar{y})^2}$$

- R^2 takes on any value between zero and one.
 - $R^2 = 1$: Perfect match between the line and the data points.
 - $R^2 = 0$: There is no linear relationship between x and y

2.0 Regression analysis

Coefficient of determination, R^2

The plot with high R^2 tend to have points closer to the regression line → models that fit the data well will have R^2 near 1 (1 = perfect fit)



2.0 Regression analysis

Interpreting R^2

- $R^2 = 1$
 - 100% of the variation in Y is explained by the variation in X, or
 - 100% of the variation in Y is explained by the regression line/model
- $R^2 = 0.72$
 - 72% of the variation in Y is explained by the variation in X, or
 - 72% of the variation in Y is explained by the regression line/model

1.0 Regression and correlation analysis
Introduction
2.0 Regression analysis
Regression model, equation and plot
Coefficient of determination, r^2
Simple linear regression
Example
3.0 Correlation analysis
Relationship between variables
Correlation coefficient, r
4.0 Summary Regression vs. correlation

2.0 Regression analysis

Regression coefficients, a and b: Manual calculation

Coefficient formula

$$\text{Slope, } b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$\text{Intercept, } a = \bar{Y} - b\bar{X}$$

Example: Age vs. blood pressure

The following are the age (in years) and systolic blood pressure of 20 healthy adults.

Age (x)	B.P (y)	Age (x)	B.P (y)
20	120	46	128
43	128	53	136
63	141	60	146
26	126	20	124
53	134	63	143
31	128	43	130
58	136	26	124
46	132	19	121
58	140	31	126
70	144	23	123

2.0 Regression analysis

Regression coefficients, a and b: Manual calculation

Table:

Add column for of x^2 and xy values,
calculated using Excel

$$\sum x = 852$$

$$\sum xy = 114466$$

$$\sum y = 2630$$

$$\sum x^2 = 41678$$

Adult	x (age)	y (BP)	xy	x^2
1	20	120	2400	400
2	43	128	5504	1849
3	63	141	8883	3969
4	26	126	3276	676
5	53	134	7102	2809
6	31	128	3968	961
7	58	136	7888	3364
8	46	132	6072	2116
9	58	140	8120	3364
10	70	144	10080	4900
11	46	128	5888	2116
12	53	136	7208	2809
13	60	146	8760	3600
14	20	124	2480	400
15	63	143	9009	3969
16	43	130	5590	1849
17	26	124	3224	676
18	19	121	2299	361
19	31	126	3906	961
20	23	123	2829	529
Total	852	2630	114486	41678

2.0 Regression analysis

Regression coefficients, a and b: Manual calculation

$$\text{Slope, } b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{114486 - \frac{852 \times 2630}{20}}{41678 - \frac{852^2}{20}} = 0.4547$$

$$\bar{X} = \frac{852}{20} = 42.6$$

$$\bar{Y} = \frac{2630}{20} = 131.5$$

$$\text{Intercept, } a = \bar{Y} - b\bar{X}$$

$$\begin{aligned} a &= 131.5 - 0.4547 \times 42.6 \\ &= 112.13 \end{aligned}$$

Regression equation,

$$\hat{y} = 112.13 + 0.4547x$$

Prediction for age = 25:

$$\text{B.P} = 112.13 + 0.4547(25)$$

$$= 123.49 = \underline{123.5 \text{ mmHg}}$$

2.0 Regression analysis | using R

Model fitting - R codes

Regression analysis can be conducted using two quantitative variables (i.e. age vs. blood pressure)

The output provides the coefficients of the **regression equation** and the **R²** for the model fitted

R codes:

To set working directory and read data from an Excel file using the “readxl” package;

```
setwd("G:/My Drive/1.TEACHING/A-SEM 1 2021_2022 (AGR3701_AGR5201)/AGR5201/R analysis")
```

```
library(readxl)
```

```
reg_bp <- read_excel("regr_age_bp.xlsx")
```

To fit the regression model:

```
fit_bp <- lm(bp ~ age, data = reg_bp)
```

To display the output

```
summary(fit_bp)
```

To check for normality assumption:

```
par(mfrow = c(2,2))
```

```
plot(bp)
```


2.0 Regression analysis | using R

Model fitting | R output and equation

R output

```
> fit_bp <- lm(bp ~ age, data = reg_bp)
> summary(fit_bp)
```

```
Call:
lm(formula = bp ~ age, data = reg_bp)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-5.0463 -1.3369  0.0442  1.5661  6.5868
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 112.12629    1.61783   69.31  < 2e-16 ***
age          0.45478     0.03544   12.83  1.7e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.6 on 18 degrees of freedom
Multiple R-squared:  0.9015, Adjusted R-squared:  0.896
F-statistic: 164.7 on 1 and 18 DF, p-value: 1.702e-10
```

Regression equation from R output:

The equation is developed using the 'the coefficient values under the 'Estimate' in R output

The diagram shows the regression equation $Y = 112.13 + 0.45 (X)$. Callouts identify the components: 'BP' points to 'Y', 'Age' points to 'X', 'Slope' points to '0.45', and 'Intercept' points to '112.13'.

$$R^2 = 0.896$$

The regression model explains 89.6% of the data variability

This value is taken from adjusted R^2 in R output

2.0 Regression analysis | using R

Regression plot, line and equation | Base R package

#To plot the regression plot, use:

```
plot (bp~age, data = reg_bp, main="Regression for age on blood pressure",  
      xlab="Age (year)", ylab="Blood pressure (mmHg)")
```

#Draw a regression line

```
abline(fit_bp, col="blue")
```



2.0 Regression analysis | using R

Regression plot | "ggpubr" package - alternative package

R codes

```
install.packages ("ggpubr")
```

```
library (ggpubr) # load the library
```

The dataset name

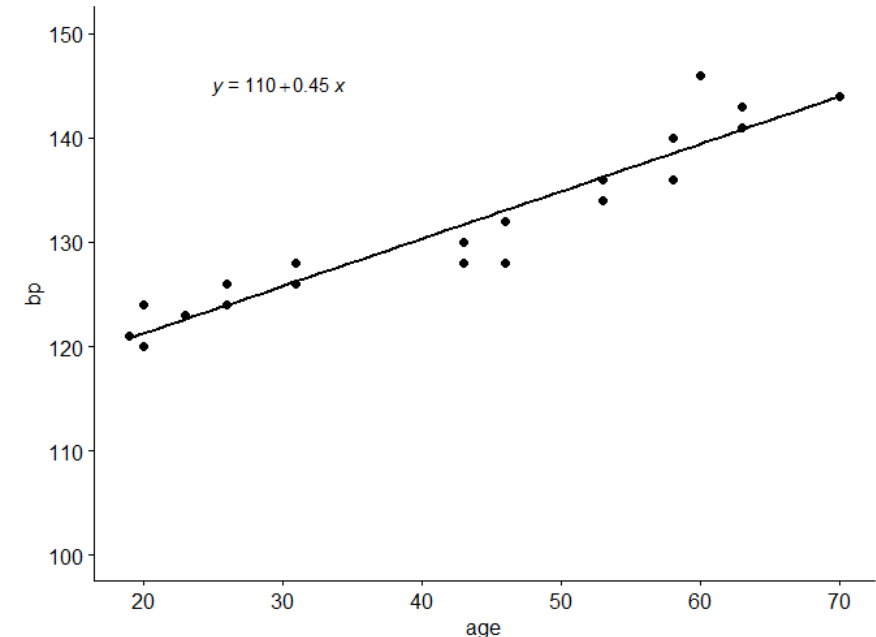
The variable name for
regression, x and y

```
ggscatter(reg_bp, x = "age", y = "bp",  
          add = "reg.line",  
          ylim = c(100,150)) +  
  stat_regline_equation(label.x = 25, label.y = 145)
```

ylim →
Set the
limit of y
axis from
100 to 150
mmHg

label.x and **label.y** → Set the
location of regression equation in
the graph at x = 25 and y = 145

R output - plot

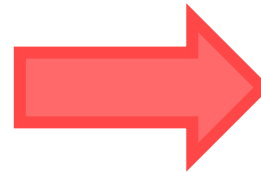
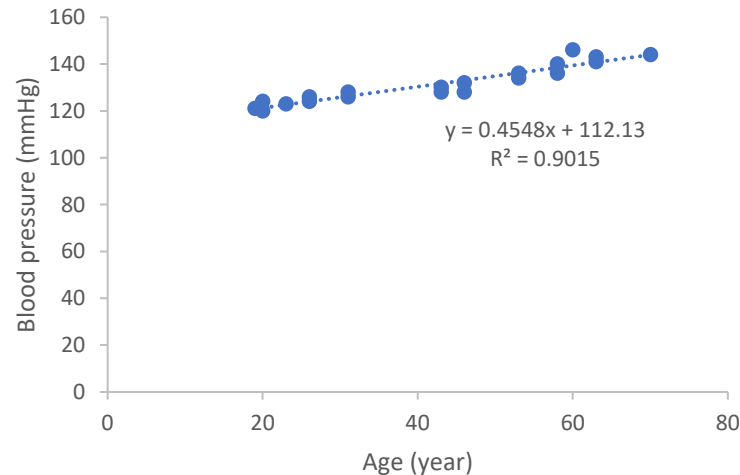


2.0 Regression analysis | using Excel

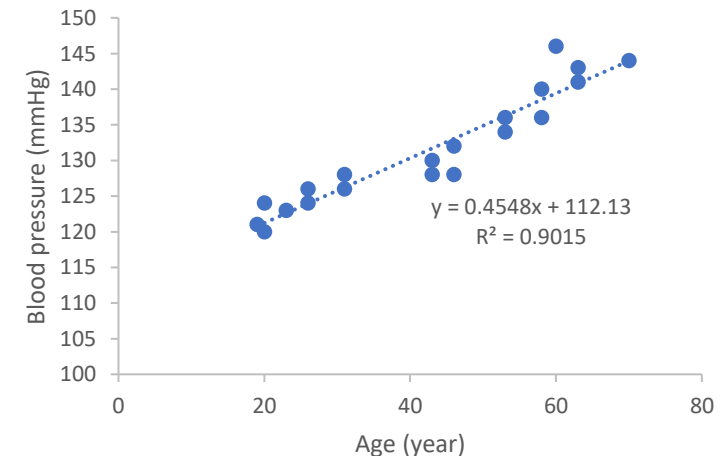
Regression plot in Excel

- To plot regression, highlight the column for **age** and **bp**, then go to **Insert > Chart**, and select **Scatter**. Then, the plot will be displayed.
- The appearance of the chart can be changed using **"Chart design" > Quick layout**. Select the preset layout that has the axis title and equation.
- Further explanations on plotting a regression plot can be viewed here → [Regression plot using Excel](#)

Original plot



Final plot after adjustment of Y axis minimum value





AGR5201

ADVANCED STATISTICAL METHODS

Regression and Correlation Analysis

Topic outline

1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

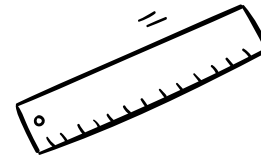
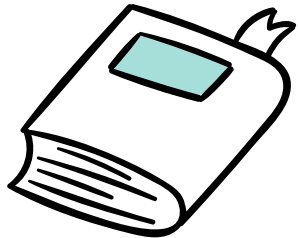
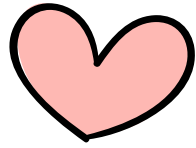
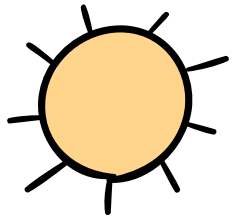
Covariance

Correlation coefficient, r

4.0 Summary | Regression vs. correlation

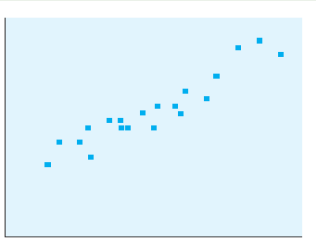
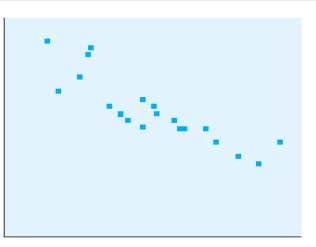
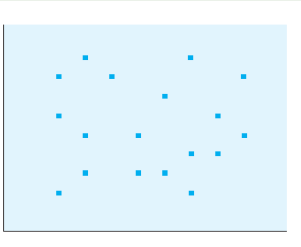
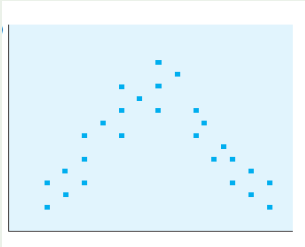


CORRELATION ANALYSIS



3.0 Correlation analysis

Relationship between two variables, X and Y

Types of relationship	POSITIVE	NEGATIVE	NONE	CURVILINEAR
Description of relationship	Both variables increase and decrease together	One variable increases, another variable decreases	Variables are not related, no trend	Variables increases together, but at one point one variable decrease as another variable increases
Description of scatterplot	Data is linear from lower left to upper right	Data is linear from upper left to lower right	No pattern - scattered all over	Form a curve either U or inverted U shape
Example	Smoking and cancer	Mountain elevation and temperature	Shoe size and IQ	Memory and age
Relationship pattern (graph)				

3.0 Correlation analysis

Relationship between two variables, X and Y

- We wish to measure the **strength** and **direction** of relationship between explanatory (X) and response (Y) variables.
 - Direction: positive or negative
 - Strength: strong or weak
- The **direction** of relationship between the two variables can be measured using:
 - i. covariance (Cov (Y, X))
 - ii. correlation coefficient (r)



What is the difference between covariance & correlation coefficient?

Covariance (Y, X)

$$\text{Cov} (Y, X) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{n - 1}$$

Correlation, r

$$\text{Cor} (Y, X) = \frac{\text{Cov} (Y, X)}{s_y s_x}$$

Where:

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}} \text{ and } s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Equivalent formula:

$$\text{Cor} (Y, X) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}}$$

3.0 Correlation analysis

1. Covariance

- Covariance indicates the direction (positive or negative) of the linear relationship between Y and X.
- If the value of:
 - $\text{Cov}(Y, X) > 0 \rightarrow$ Positive relationship
 - $\text{Cov}(Y, X) < 0 \rightarrow$ Negative relationship
- However, covariance does not tell us much about the strength of such relationship because it is affected by changes in the unit of measurement.
- To avoid this disadvantage, we standardize the X and Y data by subtract the mean from each observation and divide by its standard deviation, where:

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}} \text{ and } s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

- The covariance between standardized X and Y is known as correlation coefficient



Covariance Clearly Explained!

Click link below or scan the QR code on the right:

<https://youtu.be/TPcAnExkWwQ>



Covariance (Y, X)

$$\text{Cov}(Y, X) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{n - 1}$$

Correlation, r

$$\text{Cor}(Y, X) = \frac{\text{Cov}(Y, X)}{s_y s_x}$$

Equivalent formula:

$$\text{Cor}(Y, X) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}}$$

3.0 Correlation analysis

2. Correlation coefficient, r

- Correlation coefficient or simple correlation coefficient is a statistics showing the degree of relation between two variables:
 - It is a covariance of '*standardized X and Y*' variables
 - It is also called Pearson's correlation or product moment correlation coefficient.
 - It measures the nature and strength between two quantitative variables.

Formula of correlation coefficient, r

$$\text{Cor}(Y, X) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}}$$

Or equivalent formula



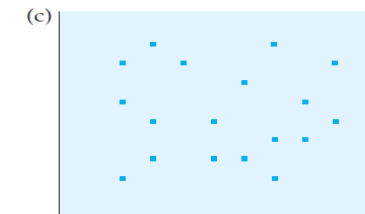
$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$



Positive linear relationship



Negative linear relationship



No relationship



Nonlinear relationship (curvilinear)

3.0 Correlation analysis

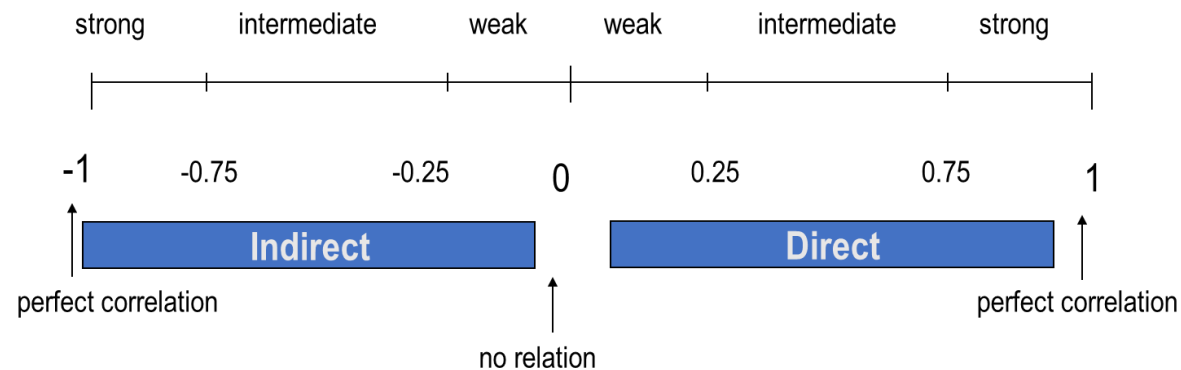
2. Correlation coefficient, r

- The sign of r denotes the nature of association, while the value of r denotes the strength of association.
- If the sign is +ve \rightarrow the relation is direct
 - Positive relationship
 - An increase in one variable is associated with an increase in the other variable.
 - A decrease in one variable is associated with a decrease in the other variable.
- If the sign is -ve \rightarrow the relationship is indirect
 - Negative relationship
 - an increase in one variable is associated with a decrease in the other

3.0 Correlation analysis

2. Correlation coefficient, r

- The value of r ranges between (-1) and (+1)
- The value of r denotes the strength of the association as illustrated by the following diagram.



Interpretations of correlation coefficients based on three different fields summarized by Akoglu (2018).

Table 1

Interpretation of the Pearson's and Spearman's correlation coefficients.

Correlation Coefficient	Dancey & Reidy (Psychology)	Quinnipiac University (Politics)	Chan YH (Medicine)
+ 1	- 1	Perfect	Perfect
+ 0.9	- 0.9	Strong	Very Strong
+ 0.8	- 0.8	Strong	Very Strong
+ 0.7	- 0.7	Strong	Very Strong
+ 0.6	- 0.6	Moderate	Moderate
+ 0.5	- 0.5	Moderate	Fair
+ 0.4	- 0.4	Moderate	Fair
+ 0.3	- 0.3	Weak	Fair
+ 0.2	- 0.2	Weak	Poor
+ 0.1	- 0.1	Weak	Poor
0	0	Zero	None

The naming on the 1) Left: Dancey & Reidy.,⁴ 2) Middle: The Political Science Department at Quinnipiac University, 3) Right: Chan et al.⁵.

Scan the QR codes, or click the link on the right for **reference**:



Akoglu (2018)
<https://tinyurl.com/agr5201>

3.0 Correlation analysis

Interpretation of r value

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

- If $r = 0 \rightarrow$ no association or correlation between the two variables.
- If $0 < r < 0.25 =$ weak correlation.
- If $0.25 \leq r < 0.75 =$ intermediate correlation.
- If $0.75 \leq r < 1 =$ strong correlation.
- If $r = 1 =$ perfect correlation.

1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

Correlation coefficient, r

4.0 Summary | Regression vs. correlation

3.0 Correlation analysis

Interpretation of correlation

- Correlation is a measure of strength of a relationship, and **DOES NOT** infer CAUSATION
- This means that high correlation only inform that the relationship is strong.
- A positive correlation infer that an increase in the first variable would correspond to the increase in the second variable - direct relationship
- For example, high (+) correlation between height and weight shows that as the height increases, the weight also increases. However, this is NOT inferring that weight is causing the increasing in height, or vice versa..

3.0 Correlation analysis

Correlation coefficient, r calculation

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

$$r = \frac{114486 - \frac{852 \times 2630}{20}}{\sqrt{\left(41678 - \frac{(852)^2}{20}\right) \left(347080 - \frac{2630^2}{20}\right)}}$$

$$r = 0.9495$$

Interpretation:

The correlation between age and BP is strong and positive.

Adult	x (age)	y (BP)	xy	x ²	y ²
1	20	120	2400	400	14400
2	43	128	5504	1849	16384
3	63	141	8883	3969	19881
4	26	126	3276	676	15876
5	53	134	7102	2809	17956
6	31	128	3968	961	16384
7	58	136	7888	3364	18496
8	46	132	6072	2116	17424
9	58	140	8120	3364	19600
10	70	144	10080	4900	20736
11	46	128	5888	2116	16384
12	53	136	7208	2809	18496
13	60	146	8760	3600	21316
14	20	124	2480	400	15376
15	63	143	9009	3969	20449
16	43	130	5590	1849	16900
17	26	124	3224	676	15376
18	19	121	2299	361	14641
19	31	126	3906	961	15876
20	23	123	2829	529	15129
Total	852	2630	114486	41678	347080

3.0 Correlation analysis

R codes

```
#correlation and correlation test
```

```
cor(x=reg_bp$age, y=reg_bp$bp, method = "pearson")
```

```
cor.test(x=reg_bp$age, y=reg_bp$bp, method = "pearson")
```

3.0 Correlation analysis

R Output

```
> #correlation and correlation test  
> cor(x=reg_bp$age, y=reg_bp$bp, method = "pearson")  
[1] 0.9494537  
> cor.test(x=reg_bp$age, y=reg_bp$bp, method = "pearson")
```

Pearson's product-moment correlation

```
data: reg_bp$age and reg_bp$bp  
t = 12.832, df = 18, p-value = 1.702e-10  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
 0.8742530 0.9801581  
sample estimates:  
cor  
0.9494537
```

The p-value for hypothesis test to test if the relationship is significant from 0 or not.

This is the correlation coefficient from R

4.0 Summary | Regression vs. Correlation

Regression

- Regression is *used* for:
 1. Develop a mathematical equation for the relationship between two quantitative variables, X and Y.
 2. Predict the value of Y based on a value of X
 3. R^2 – coefficient of determination → how much the variation in Y is explained by the regression line.

Correlation

- Determine the strength and direction of relationship between two quantitative variables.
- Correlation coefficient (r) → range from -1 to +1.
 - Sign (+ or -) direction (positive or negative relationship)
 - Value of r → strength (weak (0) or strong (1))

1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

Correlation coefficient, r

4.0 Summary | Regression vs. correlation

4.0 Summary

Regression after ANOVA

- Regression (SLR) → to study the relationship between two variables (Model I and II)
- Regression can be done after ANOVA to further investigate the relationship between response variable (DV) and quantitative treatments (IV).
- Example:
 - Rate of N? (0, 100, 150 kg /ha) → yes, it is a quantitative variable
 - Day of storage (0, 1, 2, 3 days) → quantitative
 - Variety? → No, it is a categorical variable.

1.0 Regression and correlation analysis

Introduction

2.0 Regression analysis

Regression model, equation and plot

Coefficient of determination, r^2

Simple linear regression

Example

3.0 Correlation analysis

Relationship between variables

Correlation coefficient, r

4.0 Summary | Regression vs. correlation

4.0 Summary

Exercise question | Soil moisture vs. plant growth

An experiment was conducted to study the relationship between soil moisture and plant growth. The data are given in the table on the right. The soil moisture and growth rate are both quantitative and continuous data (numerical)

1. Write the regression equation to explain the relationship between moisture and plant growth. Interpret the coefficients
2. Create a regression plot and display the equation and r^2 values in the plot.
3. Conduct a correlation test on the relationship between these two variables. Interpret the output



Sample	Moisture	Growth rate (mm/week)
1	0.35	22.86
2	1.48	35.93
3	1.71	43.45
4	0.57	28.97
5	0.42	24.49
6	0.78	31.86
7	0.70	23.97
8	0.78	28.61
9	1.64	40.14
10	1.45	37.57
11	1.77	40.42
12	0.57	21.52
13	1.17	36.67
14	1.86	49.34
15	2.09	47.26
16	0.96	37.27
17	0.45	28.57
18	1.58	35.94
19	0.39	25.03
20	1.49	39.13
21	0.96	32.88
22	2.08	42.35
23	2.06	50.14
24	0.72	22.53
25	1.16	40.53
26	1.80	39.19
27	2.09	37.52
28	1.49	32.89
29	1.56	47.07
30	2.01	37.00

THANKS!

