Introduction to biostatistics and machine learning

Olga Dethlefsen, Eva Freyhult, Bengt Sennblad, Payam Emami

2020-11-09

Contents

Preface

This "bookdown" book contains teaching and learning materials prepared and used during "Introduction to biostatistics and machine learning" course organized by NBIS, National Bioinformatics Infrastructure Sweden. The course is open for PhD students, postdoctoral researcher and other employees in need of biostatistical skills within Swedish universities. The materials are geared towards life scientists wanting to be able to understand and use basic statistical and machine learning methods. It may also suits those already applying biostatistical methods but who have never gotten a chance to reflect on the basic statistical concepts, such as the commonly misinterpreted p-value.

More about the course https://nbisweden.github.io/workshop-mlbiostatistics/

Part I

Preliminary Mathematics

Chapter 1

Mathematical notations

Aims

- to recapitulate the basic notations and conventions used in mathematics and statistics

Learning outcomes

- to recognize natural numbers, integrals and real numbers
- to understand the differences between variables and constants
- to use symbols, especially Sigma and product notations, to represent mathematical operations

1.1 Numbers

- **Natural numbers, N**: numbers such as 0, 1, 3, …
- **Integers, Z**: include negative numbers …, -2, -1, 0, 1, 2
- **Rational numbers**: numbers that can be expressed as a ratio two integers, i.e. in a form $\frac{a}{b}$, where a and b are integers, and $b \neq 0$
- **Real numbers, R**: include both rational and irrational numbers
- **Reciprocal** of any number is found by diving 1 by the number, e.g. reciprocal of 5 is $\frac{1}{5}$
- **Absolute value** of a number can be viewed as its distance from zero, e.g. the absolute value of 6 is 6, written as $|6| = 6$ and absolute value of -5 is 5, written as $|-5| = 5$
- **Factorial** of a non-negative integer number n is denoted by $n!$ and it is a product of all positive integers less than or equal to n, e.g. $4! = 4{\cdot}3{\cdot}2{\cdot}1 = 24$

1.2 Variables, constants and letters

Mathematics gives us a precise language to communicate different concepts and ideas. To be able to use it it is essential to learn symbols and understand how they are used to represent physical quantities as well as understand the rules and conventions that have been developed to manipulate them.

- **variables**: things that can vary, e.g. temperature and time
- **constants**: fixed and unchanging quantities used in certain calculations, e.g. 3.14159
- in principle one could freely choose letters and symbols to represent variables and constants, but it is helpful and choose letters and symbols that have meaning in a particular context. Hence, we
- x, y, z, the end of the alphabet is reserved for variables
- a, b, c, the beginning of the alphabet is used to represent constants
- π, ω and Greek letters below are used frequently used to represent common constant, e.g. $\pi = 3.14159$

Table 1.1: Uppercase and lowercase letters of the Greek alphabet

Letter	Upper case	Lower case	Letter	Upper case	Lower case
Alpha	A	α	Nu	N	ν
Beta	B	β	Xi	Ξ	ξ
Gamma	Γ	γ	Omicron	O	o
Delta	Δ	δ	Pi	Π	π
Epsilon	E	ϵ	Rho	P	ρ
Zeta	Z	ζ	Sigma	Σ	σ
Eta	H	η	Tau	T	τ
Theta	Θ	θ	Upsilon	Y	υ
Iota	i	ι	Phi	Φ	ϕ
Kappa	K	κ	Chi	Γ	γ
Lambda	Γ	γ	Psi	Ψ	ψ
Mu	M	μ	Omega	Ω	ω

1.3 A precise language

- Mathematics is a precise language meaning that a special attention has to be paid to the exact position of any symbol in relation to other.
- Given two symbols x and y, xy and x^y and x_y can mean different things
- xy stands for multiplication, x^y for superscript and x_y for subscript

1.4 Using symbols

If the letters x and y represent two numbers, then:

- their **sum** is written as $x + y$
- subtracting y from x is $x - y$, known also as **difference**
- to multiply x and y we written as $x \cdot y$ or also with the multiplication signed omitted as xy. The quantity is known as **product of x and y**
- multiplication is **associative**, e.g. when we multiply three numbers together, $x \cdot y \cdot z$, the order of multiplication does not matter, so $x \cdot y \cdot z$ is the same as $z \cdot x \cdot y$ or $ycdotz \cdot x$
- division is denoted by $\frac{x}{y}$ and mans that x is divided by y. In this expression x, on the top, is called **numerator** and y, on the bottom, is called **denominator**
- division by 1 leaves any number unchanged, e.g. $\frac{x}{1} = x$ and division by 0 is not allowed

Equal sign

- the equal sign = is used in **equations**, e.g. $x - 5 = 0$ or $5x = 1$
- the equal sign = can be also used in **formulae**. Physical quantities are related through a formula in many fields, e.g. the formula $A = \pi r^2$ relates circle area A to its radius r and the formula $s = \frac{d}{t}$ defines speed as distance d divided by time t
- the equal sign = is also used in identities, expressions true for all values of the variable, e.g. $(x - 1)(x - 1) = (x^2 - 1)$
- opposite to the equal sign is "is not equal to" sign $\neq$, e.g. we can write $1 + 2 \neq 4$

Sigma and Product notation

- the Σ notation, read as **Sigma notation**, provides a convenient way of writing longs sums, e.g. the sum of $x_1 + x_2 + x_3 + ... + x_{20}$ is written as $\sum_{i=1}^{i=20} x_i$
- the Π notation, read as **Product notation**, provides a convenient way of writing longs products, e.g. $x_1 \cdot x_2 \cdot x_3 \cdot ... \cdot x_{20}$ is written as $\prod_{i=1}^{i=20} x_i$

1.5 Inequalities

Given any two real numbers a and b there are three mutually exclusive possibilities:

- $a > b$, meaning that a is greater than b
- $a < b$, meaning that a is less than b
- $a = b$, meaning that a is equal to b

Strict and weak

- inequality in $a > b$ and $a < b$ is **strict**
- as oppose to **weak** inequality denoted as $a \geq b$ or $a \leq b$

Some useful relations are:

- if $a > b$ and $b > c$ then $a > c$
- if $a > b$ then $a + c > b$ for any c
- if $a > b$ then $ac > bc$ for any positive c
- if $a > b$ then $ac < bc$ for any negative c

1.6 Indices and powers

- **Indices**, also known as **powers** are convenient when we multiply a number by itself several times
- e.g. $5 \cdot 5 \cdot 5$ is written as 5^3 and $4 \cdot 4 \cdot 4 \cdot 4 \cdot 4$ is written as 4^5
- in the expression x^y, x is called the *base* and y is called *index* or *power*

The laws of indices state:

- $a^m \cdot a^n = a^{m+n}$
- $\frac{a^m}{a^n} = a^{m-n}$
- $(a^m)^n = a^{m \cdot n}$

Rules derived from the laws of indices:

- $a^0 = 1$
- $a^1 = a$

Negative and fractional indices:

- $a^{-m} = \frac{1}{a^m}$ e.g. $5^{-2} = \frac{1}{5^2} = \frac{1}{25}$ for negative indices
- e.g. $4^{\frac{1}{2}} = \sqrt{4}$ or $8^{\frac{1}{3}} = \sqrt[3]{8}$ for fractional indices

---

1.7 Exercises: notations

Exercise 1.1. Classify numbers as natural, integers or real. If reall, specify if they are rational or irrational.

a) $\frac{1}{3}$
b) 2
c) $\sqrt{4}$
d) 2.3
e) π
f) $\sqrt{5}$
g) -7
h) 0
i) 0.25

Exercise 1.2. Classify below descriptors as variables or constants. Do you know the letters or symbols commonly used to represent these?

a) speed of light in vacuum
b) mass of an apple
c) volume of an apple
d) concentration of vitamin C in an apple
e) distance from Stockholm central station to Uppsala central station
f) time on the train to travel between the above stations
g) electron charge

Exercise 1.3. Write out explicitly what is meant by the following:

a) $\sum_{i=1}^{i=6} k_i$

b) $\prod_{i=1}^{i=6} k_i$

c) $\sum_{i=1}^{i=6} i^k$

d) $\prod_{i=1}^{i=3} i^k$

e) $\sum_{i=1}^{n} i$

f) $\sum_{i=1}^{i=4} (i+1)^k$

g) $\prod_{i=1}^{i=4} (k+1)^i$

h) $\prod_{i=0}^{n} i$

Exercise 1.4. Use Sigma or Product notation to represent the long sums and products below:

a) $1+2+3+4+5+6$
b) $2^2+3^2+4^2+5^2$
c) $4 \cdot 5 \cdot 6 \cdot 7 \cdot 8$
d) $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+...+\frac{1}{n}$
e) $2-2^2+2^3-2^4+...+2^n$
f) $3+6+9+12+\cdots+60$
g) $3x+6x^2+9x^3+12x^4+...+60x^{20}$
h) $3x \cdot 6x^2 \cdot 9x^3 \cdot 12x^4 \cdot ... \cdot 60x^{20}$

Answers to selected exercises (notations)

Exr. 1.1

a) real, rational
b) natural and integers, integers include natural numbers
c) $\sqrt{4} = 2$ so it is a natural number and/subset of integers
d) real number, rational as it could be written as $\frac{23}{10}$
e) real number, irrational as it cannot be explained by a simple fraction
f) real number, irrational as it cannot be explained by a simple fraction
g) integer, non a natural number as these do not include negative numbers
h) natural number, although there is some argument about it as some define natural numbers as positive integers starting from 1, 2 etc. while others include 0.
i) real, rational number, could be written as $\frac{25}{100}$

Exr. 1.2

a) constant, speed of light in vacuum is a constant, denoted c with $c = 299792458\frac{m}{s}$
b) variable, mass of an apple is a variable, different for different apple sizes, for instance 138 grams, denoted as $m = 100g$
c) variable, like mass volume can be different from apple to apple, denoted as V, e.g. $V = 200cm^3$
d) variable, like volume and mass can vary, denoted as ρ_i and defined as $\rho_i = \frac{m}{V}$. So given 6.3 milligrams of vitamin C in our example apple, we have $\rho_i = \frac{0.0063}{2}\frac{g}{cm^3} = 0.0000315\frac{g}{cm^3}$ concentration of vitamin D
e) constant, the distance between Stockholm and Uppsala is fixed; it could be a variable though if we were to consider an experiment on a very long time scale; distance is often denoted in physics as d
f) variable, time on the train to travel between the stations varies, often denoted as t with speed being calculated as $s = \frac{d}{t}$
g) constant, electron charge is $e = 1.60217663 \cdot 10^{-19}C$

Exr. 1.3

a) $\sum_{i=1}^{i=6} k_i = k_1 + k_2 + k_3 + k_4 + k_5 + k_6$

b) $\prod_{i=1}^{i=6} k_i = k_1 \cdot k_2 \cdot k_3 \cdot k_4 \cdot k_5 \cdot k_6$

c) $\sum_{i=1}^{i=3} i^k = 1^k + 2^k + 3^k$

d) $\prod_{i=1}^{i=3} i^k = 1^k \cdot 2^k \cdot 3^k$

e) $\sum_{i=1}^{n} i = 1 + 2 + 3 + ... + n$ we are using dots (...) to represent all the number until n. Here, thanks to Gauss we can also write $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, i.e. Gauss formula for sum of first n natural numbers

Exr. 1.4

a) $1 + 2 + 3 + 4 + 5 + 6 = \sum_{k=1}^{6} k$

b) $2^2 + 3^2 + 4^2 + 5^2 = \sum_{x=2}^{5} x^2$

c) $4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 = \prod_{x=4}^{8} x$

d) $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + ... + \frac{1}{n} = \sum_{k=1}^{n} \frac{1}{k}$

Chapter 2

Sets

Aims

- to introduce sets and basic operations on sets

Learning outcomes

- to be able to explain what a set is
- to be able to construct new sets from given sets using the basic set operations
- to be able to use Venn diagrams to shows all possible logical relations between two and three sets

2.1 Definitions

- **set**: a well-defined collection of distinct objects, e.g. $S = \{2, 4, 6\}$
- **elements**: the objects that make up the set are also known as **elements** of the set
- if x is an element of S, we say that x belongs to S and write $x \in S$ and if x is not an element of S we say that x does not belong to S and write $x \notin S$
- a set may contain **finitely** many or **infinitely** many elements

- **subset,** $\subseteq$: if every element of set A is also in B, then A is said to be a subset of B, written as $A \subseteq B$ and pronounced A is contained in B, e.g. $A \subseteq B$, when $A = \{2, 4, 6\}$ and $\$ = B = \{2, 4, 6, 8, 10\}$. Every set is a subset if itself.
- **superset**: for our outs A and B we can also say that B is a **superset** of A and write $B \supset A$

- **cardinality**: the number of elements within a set S, denoted as $|S|$

- **empty set,** $\emptyset$: is a unique set with no members, denoted by $E = \emptyset$ or $E = \{\}$. The empty set is a subset of very set.

2.2 Basic set operations

- **union of two sets,** $\cup$: two sets can be "added" together, the union of A and B, written as $A \cup B$, e.g. $\{1,2\} \cup \{2,3\} = \{1,2,3\}$ or $\{1,2,3\} \cup \{1,4,5,6\} = \{1,2,3,4,5,6\}$
- **intersection of two sets,** $\cap$: a new set can be constructed by taking members of two sets that are "in common", written as $A \cap B$, e.g. $\{1,2,3,4,5,6\} \cap \{2,3,7\} = \{2,3\}$ or $\{1,2,3\} \cap \{7\} = \{\emptyset\}$

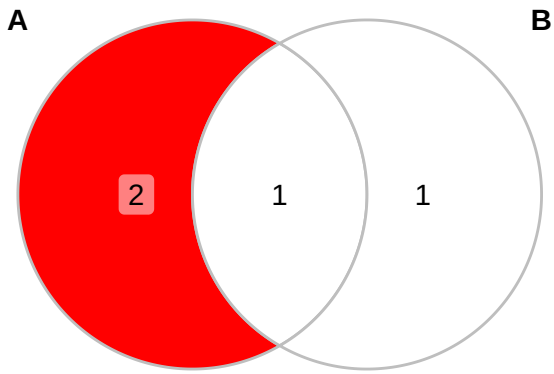
- **complement of a set,** A', A^c: are the elements not in A
- **difference of two sets,** : two sets can be "substracted", denoted by $A \ B$, by taking all elements that are members of A but are not members of B, e.g. $\{1,2,3,4\} \ \{1,3\} = \{2,4\}$. This is also in other words a relative complement of A with respect to B.

- **partition of a set**: a partition of a set S is a set of nonempty subset of S, such that every element x in S is in exactly one of these subsets. That is, the subset are pairwise *disjoint*, meaning no two sets of the partition contain elements in common, and the union of all the subset of the partition is S, e.g. Set $\{1,2,3\}$ has five partitions: i) $\{1\},\{2\},\{3\}$, ii) $\{1,2\},\{3\}$, iii) $\{1,3\},\{2\}$, iv) $\{1\},\{2,3\}$ and v) $\{1,2,3\}$

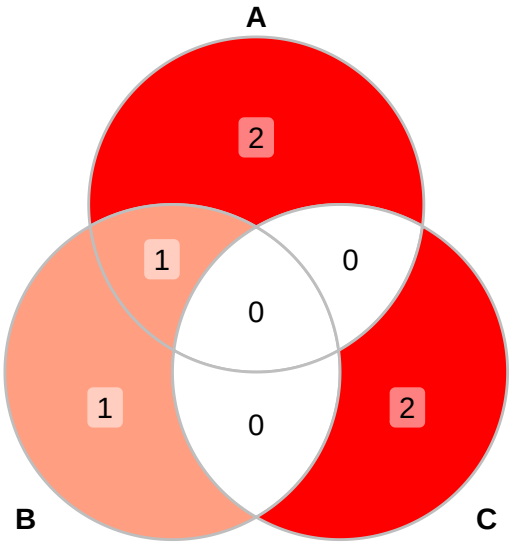
2.3 Venn diagrams

Venn diagram is a diagram that shows all possible logical relations between a finite collection of different sets. A Venn diagrams shows elements as points in the plane, and sets as regions inside closed curves. A Venn diagram consists of multiple overlapping closed curves, usually circles, each representing a set.

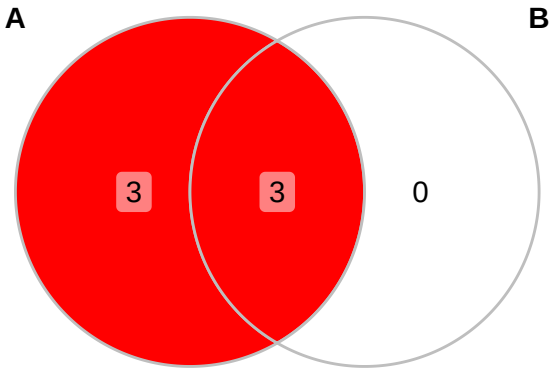
E.g. given $A = \{1,2,5\}$ and $B = \{1,6\}$ Venn diagram of A and B:



And given $A = \{1, 2, 5\}$, $B = \{1, 6\}$ and $C = \{4, 7\}$ Venn diagram of A, B and C:



And given $A = \{1, 2, 3, 4, 5, 6\}$ and $B = \{2, 4, 6\}$ Venn diagram of A and B:



2.4 Exercises: sets

Exercise 2.1. Given set $S = \{1, 2, 3, 4, 5, 6\}$:

a) what is the subset T of S consisting of its even elements?
b) what is the complement T^c?
c) what is the subset U of S containing of the prime numbers in S?
d) what is the intersection $T \cap U$?
e) what is the union of $T \cup U$?
f) what is the set difference $U \quad T$?

Exercise 2.2. Given set

$A = \{cat, elephant, dog, turtle, goldfish, hamster, parrot, tiger, guineapig, lion\}$

a) what is the subset D of A consiting of domesticated animals?
b) what is the subset C of A consiting of Felidae (cat) family?
c) what is the interection of D and C?
d) what is the complement of D, D^c?
e) what is the union of D and C?
f) what is the set difference of A C?
g) can you draw Venn diagram showing relationship between D and C?

Answers to selected exercises (sets)

Exr. 2.1

a) $T = \{2, 4, 6\}$
b) $T^c = \{1, 3, 5\}$, i.e. T^c contains all the elements of S not in T
c) $U = \{2, 3, 5\}$, the primes in S
d) $T \cap U = \{2\}$, common elements of T and U, i.e. the even and prime numbers
e) $T \cup U = \{2, 3, 4, 5, 6\}$
f) $U \quad T = \{3, 5\}$, consisting of the elements of U that are not in T

Chapter 3

Functions

Aims

- to revisit the concept of a function

Learning outcomes

- to be able to explain what function, function domain and function range are
- to be able to identify input, output, argument, independent variable, dependent variable
- to be able to evaluate function for a given value and plot the function

3.1 Definitions

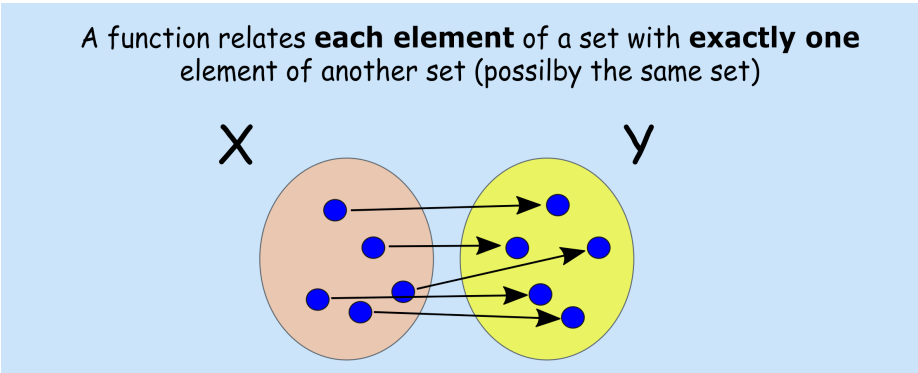


Figure 3.1: Formal function defition

- A **function**, $f(x)$, can be viewed as a rule that relates input x to an output $f(x)$
- In order for a rule to be a function it must produce a single output for any given input
- Input x is also known as **argument** of the function
- **Domain of a function**: the set of all values that the function "maps"
- **Range**: the set of all values that the function maps into

Many names are used interchangeably

Functions have been around for a while and there are many alternative names and writing conventions are being used. Common terms worth knowing:

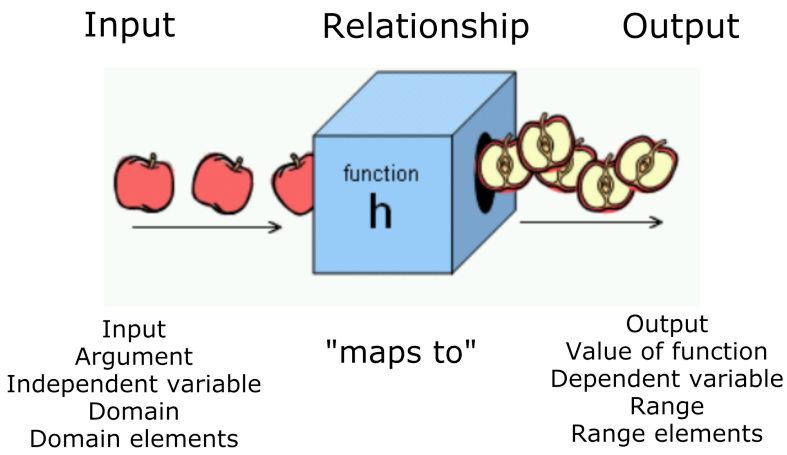


Figure 3.2: Common function terms

3.2 Evaluating function

To evaluate a function is to replace (substitute) its variable with a given number or expression. E.g. given a rule (function) that maps temperature measurements from Celsius to Fahrenheit scale:

$$f(x) = 1.8x + 32$$

where x is temperature measurements in Celsius and $f(x)$ is the associated value in Fahrenheit, we can find for a given temperature in Celsius corresponding temperature in Fahrenheit. Let's say we measure 10 Celsius degrees one autumn day in Uppsala and we want to share this information with a friend in USA. We can find the equivalent temperature in Fahrenheit by evaluating our function at

10, $f(10)$, giving us

$$f(10) = 1.8 \cdot 10 + 32 = 50$$

3.3 Plotting function

Function graphs are a convenient way of showing functions, by looking at the graph it is easier to notice function's properties, e.g. for which input values functions yields positive outcomes or whether the function is increasing or decreasing. To graph a function, one can start by evaluating function at different values for the argument x obtaining $f(x)$, plotting the points by plotting the pairs $(x, f(x))$ and connecting the dots. E.g. evaluating our above temperature rule at -20, -10, 0, 10, 20, 30 Celsius degrees results in:

x (Celsius degrees)	evaluates	f(x) (Farenheit degress)
-20	$f(-20) = 1.8 \cdot (-20) + 32$	-4
-10	$f(-20) = 1.8 \cdot (-10) + 32$	14
0	$f(-20) = 1.8 \cdot (0) + 32$	32
10	$f(-20) = 1.8 \cdot (10) + 32$	50
20	$f(-20) = 1.8 \cdot (20) + 32$	68
20	$f(-20) = 1.8 \cdot (30) + 32$	86

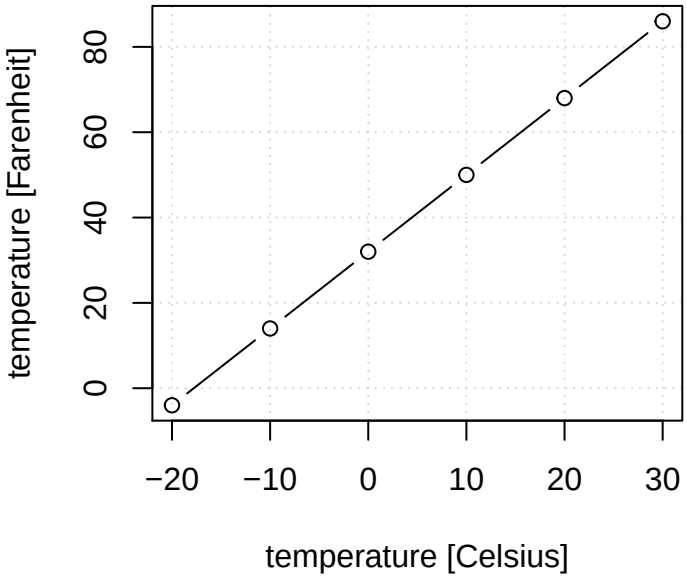


Figure 3.3: Graph of f(x) for the temeprature rule

3.4 Standard classes of functions

Algebraic function: functions that can be expressed as the solution of a polynomial equation with integer coefficients, e.g.

- constant function $f(x) = a$
- identity function $f(x) = x$
- linear function $f(x) = ax + b$
- quadratic function $f(x) = a + bx + cx^2$
- cubic function $fx() = a + bx + cx^2 + dx^3$

Transcedental functions: functions that are not algebraic, e.g.

- exponential function $f(x) = e^x$
- logarithimic function $f(x) = log(x)$
- trigonometric function $f(x) = -3sin(2x)$

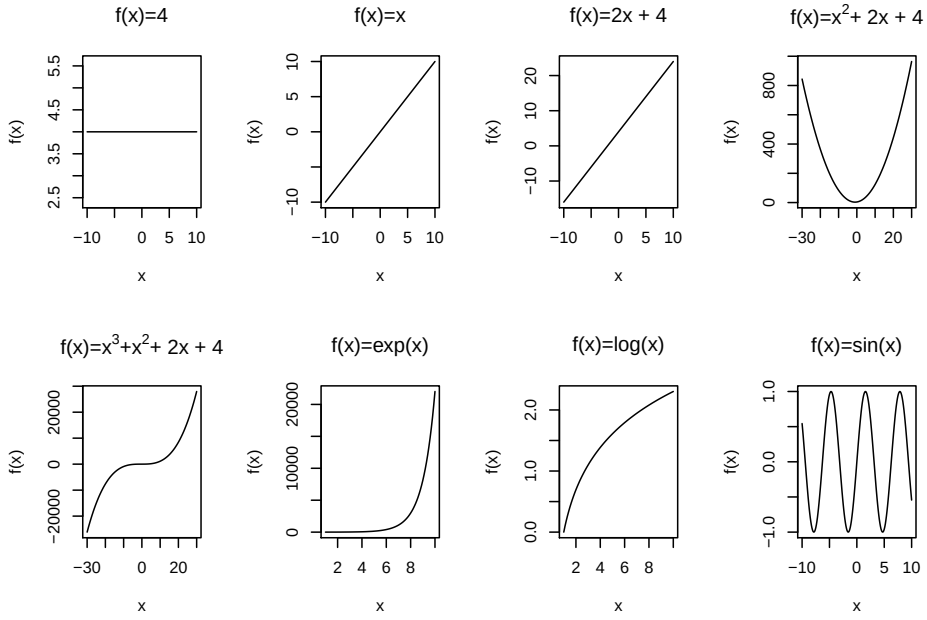


Figure 3.4: Examples of the standard classess of functions

3.5 Piecewise functions

A function can be in pieces, i.e. we can create functions that behave differently based on the input x value. They are useful to describe situations in w which a rule changes as the input value crosses certain "boundaries". E.g. a function value could be fixed in a given range and equal to the input value (identify function) for input values outside this range

$$f(x) = \begin{cases} 2 & \text{if } x \leq 1 \\ x & \text{if } x > 1 \end{cases} \tag{3.1}$$

The function can be split in many pieces, e.g. the personal training fee in SEK may depend whether the personal trainer is hired for an hour, two hours or three or more hours:

$$f(h) = \begin{cases} 500 & \text{if } h \leq 1 \\ 750 & \text{if } 1 < h \leq 2 \\ 500 + 250 \cdot h & \text{if } h > 2 \end{cases} \tag{3.2}$$

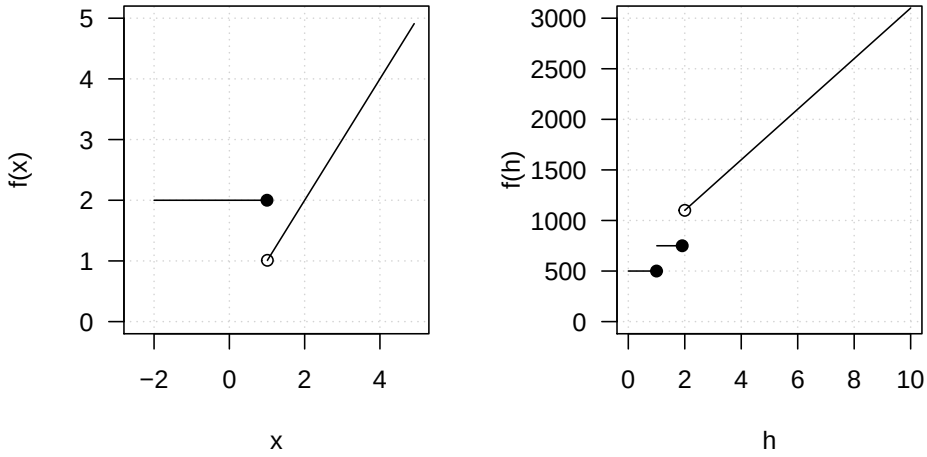


Figure 3.5: Examples of piece-wise functions

3.6 Exercises: functions

Exercise 3.1. Given the function for the personal trainer costs:

$$f(h) = \begin{cases} 500 & \text{if } h \leq 1 \\ 750 & \text{if } 1 < h \leq 2 \\ 500 + 250 \cdot h & \text{if } h > 2 \end{cases} \tag{3.3}$$

How much would you pay

a) for a 4-hours session? Evaluate function f(h) for value 4.
b) for a 2-hour session? Evalue function f(h) for value 2.

Exercise 3.2. A museum charges 50 SEK per person for a guided tour with a group of 1 to 9 people or a fixed 500 SEK fee for a group of 10 or more people. Write a function relating the number of people n to the cost C.

Exercise 3.3. Given function

$$f(x) = \begin{cases} x^2 & \text{if } x \leq 1 \\ 3 & \text{if } 1 < x \leq 2 \\ x & \text{if } x > 2 \end{cases} \tag{3.4}$$

a) sketch a graph of a function for $x \in (-4, 4)$, i.e.. for x between -4 and 4
b) evaluate function at f(1)
c) evaluate function at f(4)

Answers to selected exercises (functions)

Exr. 3.1

a) $f(4) = 500 + 250 \cdot 4 = 1500$
b) $f(2) = 750$ as $h \leq 2$ means less or equal to 2, that is including 2

Chapter 4

Differentiation

Aims

- introduce the concept of differentiation and rules of differentiation

Learning outcomes

- to be able to explain differentiation in terms of rate of change
- to be able to find derivatives in simple cases

4.1 Rate of change

- We are often interested in the rate at which some variable is changing, e.g. we may be interested in the rate at which the temperature is changing or the rate of water levels increasing
- Rapid or unusual changes may indicate that we are dealing with unusual situations, e.g. global warming or a flood
- Rates of change can be positive, negative or zero corresponding to a variable increasing, decreasing and non-changing

The function $f(x) = x^4 - 4x^3 - x^2 - e^{-x}$ changes at different rates for different values of x, e.g.

- between $x \in (-10, -9)$ the $f(x)$ is increasing at slightly higher pace than $x \in (5, 6)$
- between $x \in (-7, -5)$ the $f(x)$ is decreasing and
- between $x \in (0, 1)$ the $f(x)$ is not changing
- to be able to talk more precisely about the rate of change than just saying "large and positive" or "small and negative" change we need to quantify the changes, i.e. assign the rate of change an exact value
- **Differentiation** is a technique for calculating the rate of change of any function

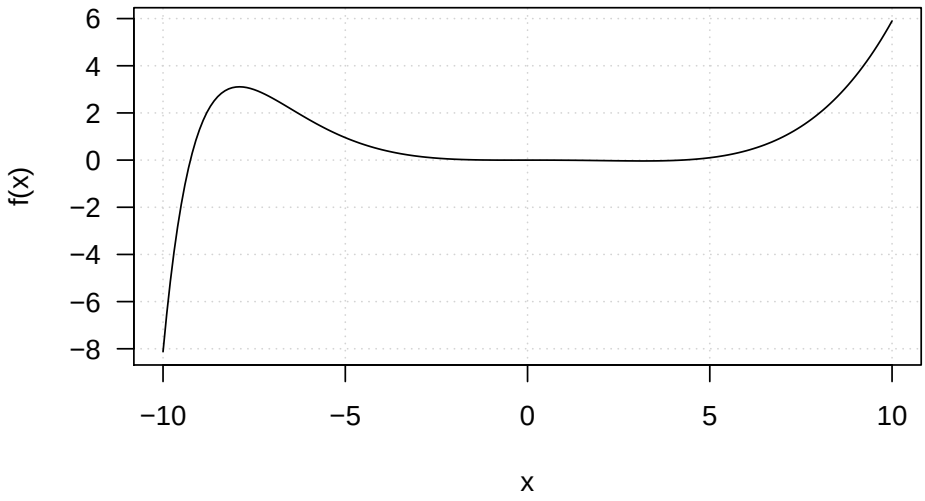


Figure 4.1: The function $f(x)$ changes at different rates for different values of x

4.2 Average rate of change across an interval

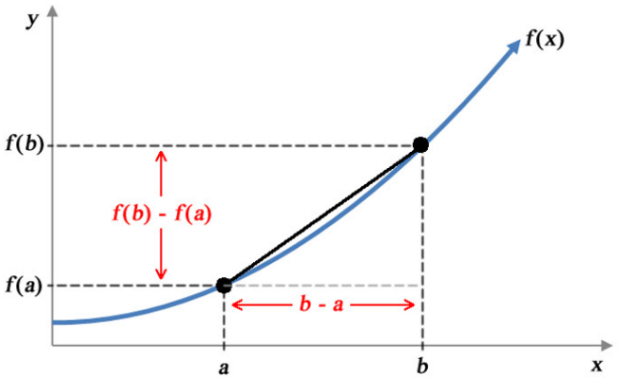


Figure 4.2: The average rate of change of $f(x)$ with respect to x over $[a, b]$ is equal to the slope of the secant line (in black)

To dive further into calculating the rate of change let's look at Figure 4.2 and define the *average rate of change* of a function across an interval. Figure 4.2 shows a function $f(x)$ with two possible argument values a and b marked and their corresponding function values $f(a)$ and $f(b)$.

Consider that x is increasing from a to b. The change in x is $b - a$, i.e. as x increases from a to b the function $f(x)$ increase from $f(a)$ to $f(b)$. The change in $f(x)$ is $f(b) - f(a)$ and the average rate of change of y across the $[a, b]$ interval

is:

$$\frac{change\ in\ y}{change\ in\ x} = \frac{f(b) - f(a)}{b - a} \tag{4.1}$$

E.g. let's take a quadratic function $f(x) = x^2$ and calculate the average rate of change across the interval $[1, 4]$.

- The change in x is $4-1$ and the change in $f(x)$ is $f(4)-f(1) = 4^2-1^2 = 16-1 = 15$. So the average rate of change is $\frac{15}{3} = 5$. What does this mean? It means that across the interval $[1, 4]$ on average the $f(x)$ value increases by 5 for every 1 unit increase in x.
- If we were to look at the average rate of change across the interval $[-2, 0]$ we would get $\frac{f(0)-f(-2)}{0-(-2)} = \frac{0-(-2)^2}{2} = \frac{-4}{2} = -2$. Here, over the $[-2, 0]$ on average the $f(x)$ value decreases by 2 for every 1 unit increase in x.
- Looking at the graph of $f(x) = x^2$ verifies our calculations

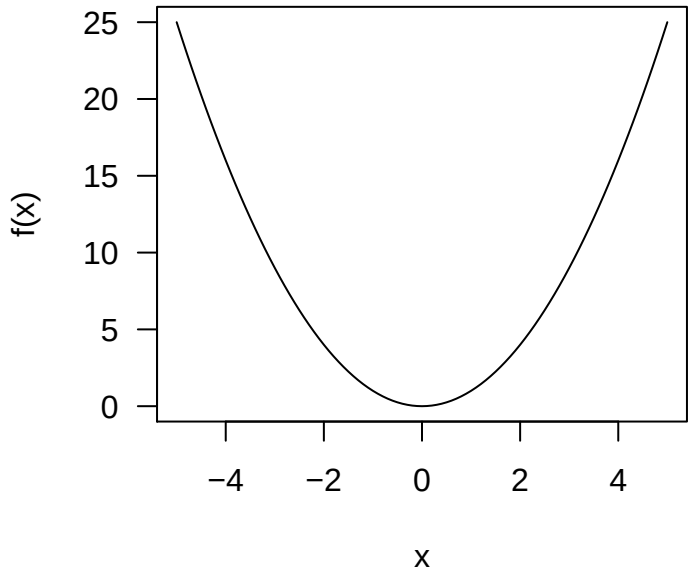


Figure 4.3: Example function $f(x) = x^2$

4.3 Rate of change at a point

- We often need to know the rate of change of a function at a point, and not simply an average rate of change across an interval.
- Figure 4.4, similar to Figure 4.2, shows, instead of two points a and b, point a and a second point defined in terms of its distance from the first

point a. Thus, the two points are now a and $a + h$ and the distance between the two points is equal to h.

- Now we can write that:

$$\frac{change\ in\ y}{change\ in\ x} = \frac{f(a+h) - f(a)}{a + h - a} = \frac{f(a+h) - f(a)}{h}$$

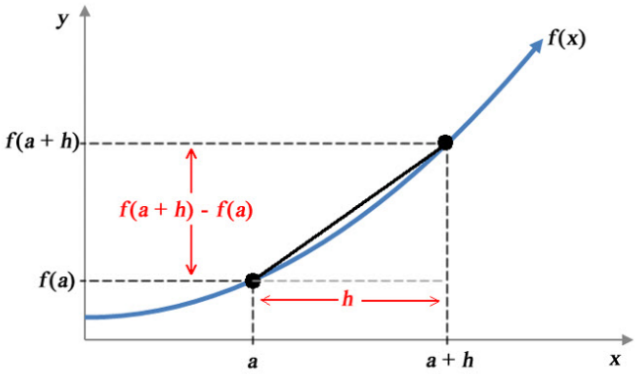


Figure 4.4: The average rate of change of $f(x)$ with respect to x over $[a, b]$ is equal to the slope of the secant line (in black)

Further:

- if we assume that the second point $a+h$ is really close to a, meaning that h approaches 0, denoted as $h \to 0$, we can find the rate of change at the point a
- the distance between the two points a and $a + h$ is getting smaller and so is the difference of the function values $f(a + h) - f(a)$. We denote these small differences as δx and δy, pronounced "delta x" and "delta y", respectively.
- the term δ reads as "delta" and represents a small change

We can thus continue and write that a **rate of change of a function at a point** is given by

$$\frac{small\ change\ in\ y}{small\ change\ in\ x} = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} \tag{4.2}$$

E.g. let's look at the linear function $f(x) = 2x + 3$. We can find the rate of change at any point of x by:

$$\frac{small\ change\ in\ y}{small\ change\ in\ x} = \frac{f(x+h) - f(x)}{x + h - x} = \lim_{h \to 0} \frac{2(x+h) + 3 - (2x+3)}{x + h - x} = \lim_{h \to 0} \frac{2h}{h} = 2$$

It means that the function value $f(x)$ increases by 2 for every small increase, h, in x. Here, this increase is the same for all the values of x, i.e. it does not

depend on x. **The change in function value** $f(x)$ **can depend** on the value of x, for instance if we look at the quadratic $f(x) = x^2$ function, we get:

$$\frac{\textit{small change in } y}{\textit{small change in } x} = \frac{f(x+h) - f(x)}{x + h - x} = \lim_{h \to 0} \frac{x^2 + 2xh + h^2 - x^2}{h} = \lim_{h \to 0} \frac{2xh + h^2}{h} = 2x + h$$

This means that:

- the rate of change for the function $f(x)$ at a point x is $2x$
- the $f(x)$ value increases by $2x$ for every small increase, h, in x
- the rate of change along a quadratic function is changing constantly according to the value of x we are looking at, it is a function of x
- and finally that the rate of change does not give us any information about the rate of change globally.

4.4 Terminology and notation

- **differentiation** is the process of finding the rate of change of a given function
- the function is said to be **differentiated**
- the rate of change of a function is also known as the **derivative** of the function
- given a function $f(x)$ we say that we differentiate function in respect to x and write:

$$\lim_{h \to 0} \frac{\delta y}{\delta x} = \frac{dy}{dx}$$

or use the "prime"

$$f'(x)$$

4.5 Table of derivatives

- in practice, there is no need to compute $\lim_{h \to 0} \frac{\delta y}{\delta x}$ every time when we want to find a derivative of a function
- instead, we can use patterns of the common functions and their derivatives

Table 4.1: Common functions and their derivatives, k denotes a constant

Function $f(x)$	Derivative $f'(x)$
k	0
x	1
kx	k
x^n	nx^{n-1}

Function $f(x)$	Derivative $f'(x)$
kx^n	knx^{n-1}
e^x	e^x
e^{kx}	ke^{kx}
$\ln(x)$	$\frac{1}{x}$
$\ln(kx)$	$\frac{1}{x}$

We can use the Table 4.1 to find derivatives of some of the functions e.g.

- $f(x) = 3x$, $f'(x) = 3$
- $f(x) = 2x^4$, $f'(x) = 2 * 4x^{4-1} = 8x^3$
- $f(x) = e^{2x}$, $f'(x) = 2e^{2x}$
- $f(x) = \ln(4x)$, $f'(x) = \frac{4}{x}$

4.6 Exercises (differentiation)

Exercise 4.1. # Find derivatives of the functions

a) $f(x) = 2$
b) $f(x) = 2x + 1$
c) $f(x) = 5x^2$
d) $f(x) = 4x^3 + x^2$
e) $f(x) = \sqrt(x)$
f) $f(x) = \ln(2x)$
g) $f(x) = e^x$
h) $f(x) = \frac{9}{x^2} + ln(4x)$
i) $f(x) = 4x - 6x^6$
j) $f(x) = \frac{3}{x^2}$

Answers to selected exercises (differentiation)

Exr. 4.1

a) $f(x) = 2$, $f'(x) = 0$
b) $f(x) = 2x + 1$, $f'(x) = 2$
c) $f(x) = 5x^2$, $f'(x) = 10x$
d) $f(x) = 4x^3 + x^2$, $f'(x) = 12x^2 + 2x$
e) $f(x) = \sqrt(x) = x^{\frac{1}{2}}$, $f'(x) = \frac{1}{2}x^{\frac{1}{2}-1} = \frac{1}{2}x^{-\frac{1}{2}}$
f) $f(x) = \ln(2x)$, $f'(x) = \frac{1}{x}$
g) $f(x) = e^x$, $f'(x) = e^x$

Chapter 5

Integration

Aims

- to introduce the concept of integration

Learning outcomes

- to be able to explain what integration is
- to be able to explain the relationship between differentiation and integration
- to be able to integrate simple functions
- to to able to use integration to calculate the area under the curve in simple cases

5.1 Reverse to differentiation

- when a function $f(x)$ is known we can differentiate it to obtain the derivative $f'(x)$
- the reverse process is to obtain $f(x)$ from the derivative
- this process is called **integration**
- apart from simple reversing differentiation integration comes very useful in finding **areas under curves**, i.e. the area above the x-axis and below the graph of $f(x)$, assuming that $f(x)$ is positive
- the symbol for integration is $\int$ and is known as "integral sign"

E.g. let's take a function $f(x) = x^2$. Suppose we only have a derivative, which is $f'(x) = 2x$ and we would like to find the function given this derivative. Formally we write:

$$\int 2x dx = x^2 + c$$

where:

- the term $2x$ within the integral is called the **integrand**
- the term dx indicates the name of the variable involved, here x
- c is **constant of integration**

5.2 What is constant of integration?

- Integration reverses the process of differentiation, here, given our example function $f(x) = x^2$ that we pretended we do not know, we started with the derivative $f'(x) = 2x$ and via integration we obtained back the very function
$$\int 2x dx = x^2$$
- However, many function can result in the very same derivative since the derivative of a constant is 0 e.g. a derivatives of $f(x) = x^2$, $f(x) = x^2 + 10$ and $f(x) = x^2 + \frac{1}{2}$ all equal to $f'(x) = 2x$
- We have to take this into account when we are integrating, i.e. reverting differentiation. As we have no way of knowing what the original function constant is, we add it in form of c, i.e. unknown constant, called the constant of integration.

5.3 Table of integrals

Similar to differentiation, in practice we can use tables of integrals to be able to find integrals in simple cases

Table 5.1: Common functions and their integrals, k denotes a constant

Function $f(x)$	Integral $\int f(x)dx$
$constant\ k$	$kx + c$
x	$\frac{x^2}{2} + c$
kx	$k\frac{x^2}{2} + c$
x^n	$\frac{x^{n+1}}{n+1} + c\ \ if\ n \neq -1$
kx^n	$k\frac{x^{n+1}}{n+1} + c$
e^x	$e^x + c$
e^{kx}	$\frac{e^{kx}}{k} + c$
$\frac{1}{x}$	$\ln(x) + c$

E.g.

- $\int 4x^3 dx = \frac{4x^{3+1}}{3+1} = x^4 + c$
- $\int (x^2 + x)dx = \frac{x^3}{3} + \frac{x^2}{2} + c$ (note: we can evaluate integrals separately and add them as integration as differentiation is linear)

5.4 Definite integrals

- the above examples of integrals are **indefinite integrals**, the result of finding an indefinite integral is usually a function plus a constant of integration
- we have also **definite integrals**, so called because the result is a definite answer, usually a number, with no constant of integration
- definite integrals are often used to areas bounded by curves or, as we will cover later on, estimating probabilities
- we write:

$$\int_a^b f(x)dx$$

where:

- $\int_a^b f(x)dx$ is called the definite integral of $f(x)$ from a to b
- the numbers a and b are known as lower and upper limits of the integral

E.g. let's look at the function $f(x) = x$ plotted below and calculate a definite integral from 0 to 2.

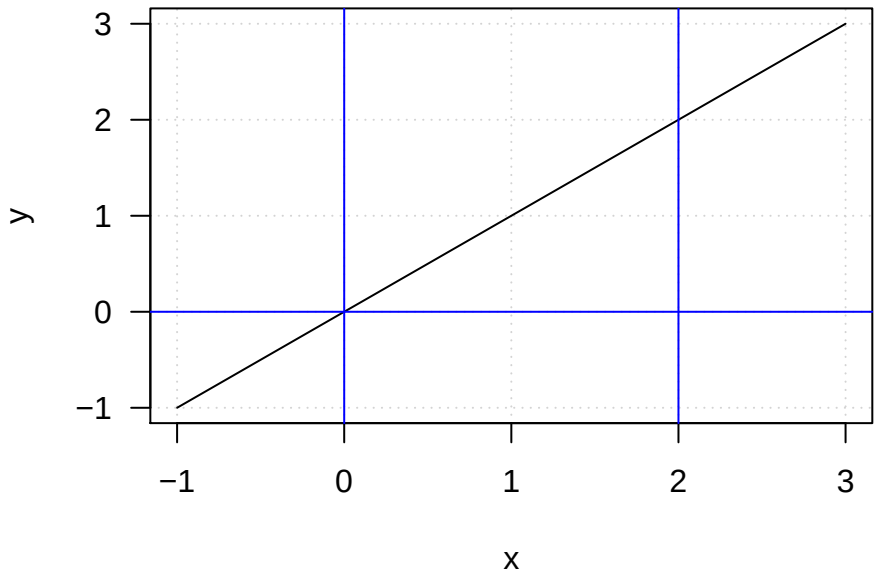


Figure 5.1: Graph of function $f(x) = x$

We write

$$\int_0^2 f(x)dx = \int_0^2 xdx = \Big[\frac{1}{2}x^2\Big]_0^2 = \frac{1}{2}(2)^2 - \frac{1}{2}(0)^2 = 2$$

so first find the integral and then we evaluate it at upper limit and subtracting the evaluation at the lower limit. Here, the result it 2. What would be the result if you tried to calculate the triangle area on the above plot, area defined by the blue vertical lines drawn at 0 and 2 and horizontal x-axis? The formula for the triangle area is $Area = \frac{1}{2} \cdot base \cdot height$ so here $Area = \frac{1}{2} \cdot 2 \cdot 2 = 2$ the same result as achieved with integration.

Exercise 5.1. Integrate:

a) $\int 2 \cdot dx$
b) $\int 2x \cdot dx$
c) $\int (x^4 + x^2 + 1) \cdot dx$
d) $\int e^x \cdot dx$
e) $\int e^{2x} \cdot dx$
f) $\int \frac{2}{x} \cdot dx$
g) $\int_2^4 2x \cdot dx$
h) $\int_0^4 (x^2 + 1) dx$
i) $\int (x^4 + \frac{2}{x} + e^{2x}) dx$
j) $\int_0^4 (x^4 + 1) dx$

Answers to selected exercises (integration)

Exr. 5.1

a) $\int 2 \cdot dx = 2x + c$
b) $\int 2x \cdot dx = \frac{2x^2}{2} = x^2 + c$
c) $\int (x^4 + x^2 + 1) \cdot dx = \frac{x^5}{5} + \frac{x^3}{3} + x + c$
d) $\int e^x \cdot dx = e^x + c$
e) $\int e^{2x} \cdot dx = \frac{1}{2} e^{2x}$
f) $\int \frac{2}{x} \cdot dx = \int 2 \cdot \frac{1}{x} \cdot dx = 2 \ln x + c$
g) $\int_2^4 2x \cdot dx = \Big[x^2\Big]_2^4 = 16 - 4 = 12$
h) $\int_0^4 (x^2 + 1) dx = \Big[\frac{x^3}{3} + x\Big]_0^4 = \frac{4^3}{3} + 4 - 0 = \frac{64}{3} + 4 = \frac{76}{3}$

Chapter 6

Vectors

Aims

- to introduce vectors and basic vectors operations

Learning outcomes

- to be able to write n-dimensional vectors using vector notations
- to be able to perform addition and scalar multiplication
- to be able to check if two vectors are orthogonal

A large number of statistical models use vectors and matrices, both for compact representations, and for the calculations, e.g. parameter estimates.

6.1 Vectors

- A vector is an ordered set of number
- These numbers, e.g. in vector $\mathbf{x}$ can be expressed as a row $\mathbf{x} = [6 \quad 0 \quad 5 \ldots 1]$
- or as a column $\mathbf{x} = \begin{bmatrix} 6 \\ 0 \\ 5 \\ \vdots \\ 1 \end{bmatrix}$
- the number of elements in a vector is referred to as its **dimension** and we often use n to express n-dimensional vector, where n can be any natural number
- here, we denote vectors using small bold font $\mathbf{x}$, other notations may include an arrow $\vec{x}$ or overline $\overline{x}$
- also **parentheses** are used interchangeably with **square bracket**,

e.g. $\mathbf{x} = [6 \quad 0 \quad 5 \ldots 1]$ can be written as $\mathbf{x} = (6 \quad 0 \quad 5 \ldots 1)$ or $\begin{pmatrix} 6 \\ 0 \\ 5 \\ \vdots \\ 1 \end{pmatrix}$

6.2 Operations on vectors

Given two vectors of the same dimension: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$

Addition: we add two vectors, element by element $\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \\ \vdots \\ x_n + y_n \end{bmatrix}$

Scalar multiplication: we can multiple vector by a numerical value, scalar, denoted as γ:

$$\gamma \cdot \mathbf{x} = \begin{bmatrix} \gamma \cdot x_1 \\ \gamma \cdot x_2 \\ \gamma \cdot x_3 \\ \vdots \\ \gamma \cdot x_n \end{bmatrix}$$

Difference $\mathbf{x} - \mathbf{y}$ can be written as $\mathbf{x} + (-1) \cdot \mathbf{y}$, thus we multiply second vector with -1 and then add two vectors

Linear combination of vectors: the vector $\gamma \cdot \mathbf{x} + \delta \cdot \mathbf{y}$ is said to be a linear combination of $\mathbf{x}$ and $\mathbf{y}$:

$$\gamma \cdot \mathbf{x} + \delta \cdot \mathbf{y} = \begin{bmatrix} \gamma \cdot x_1 + \delta \cdot y_1 \\ \gamma \cdot x_2 + \delta \cdot y_2 \\ \gamma \cdot x_3 + \delta \cdot y_3 \\ \vdots \\ \gamma \cdot x_n + \delta \cdot y_n \end{bmatrix}$$

Inner product of vectors is given by:

$$\mathbf{x} \cdot \mathbf{y} = x_1 \cdot y_1 + x_2 \cdot y_2 + \ldots x_n \cdot y_n = \sum_{i=1}^{n} x_i \cdot y_i$$

Orthogonality of vectors: two vectors are said to be orthogonal if their inner

product is zero

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i \cdot y_i = 0$$

6.3 Null and unit vector

- a **null vector** is a vector whose elements are all 0; the difference between any vector and itself yields a null vector
- a **unit vector** is a vector whose elements are all 1

Exercise 6.1. Based on vector definitions and operations:

a) find the vector $\mathbf{x} + \mathbf{y}$ when $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix}$

b) find the vector $2\mathbf{x} - \mathbf{y}$ when $\mathbf{x} = \begin{bmatrix} -2 \\ 3 \\ 5 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 0 \\ -4 \\ 7 \end{bmatrix}$

c) are $\mathbf{u}$ and $\mathbf{v}$ vectors orthogonal when when $\mathbf{u} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$?

d) are $\mathbf{u}$ and $\mathbf{v}$ vectors orthogonal when when $\mathbf{u} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 7 \\ 5 \end{bmatrix}$?

e) find the value n such that the vectors $\mathbf{u} = \begin{bmatrix} 2 \\ 4 \\ 1 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} n \\ 1 \\ 8 \end{bmatrix}$ are orthogonal.

Answers to selected exercises (vectors and matrices)

Exr. 6.1

a)

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} + \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1+0 \\ 2+3 \\ 5+1 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 6 \end{bmatrix}$$

b)

$$2\mathbf{x} - \mathbf{y} = \begin{bmatrix} 2 \cdot (-2) \\ 2 \cdot 3 \\ 2 \cdot 5 \end{bmatrix} + \begin{bmatrix} (-1) \cdot 0 \\ (-1) \cdot (-4) \\ (-1) \cdot 7 \end{bmatrix} = \begin{bmatrix} -4 \\ 6 \\ 10 \end{bmatrix} + \begin{bmatrix} 0 \\ 4 \\ -7 \end{bmatrix} = \begin{bmatrix} -4+0 \\ 6+4 \\ 10-7 \end{bmatrix} = \begin{bmatrix} -4 \\ 10 \\ 3 \end{bmatrix}$$

c) Yes, to check orthogonality we need to calculate the inner product of two vectors and see if it is equal to 0, here $\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{2} u_i \cdot v_i = 1 \cdot 2 + 2 \cdot (-1) = 2 - 2 = 0$

d) No, since the inner product does not equal to 0

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{2} u_i \cdot v_i = 3 \cdot 7 + (-1) \cdot 5 = 21 - 5 = 16 \neq 0$$

Chapter 7

Matrices

Aims

- to introduce matrix and basic matrices operations

Learning outcomes

- to be able to write matrices using matrix notations
- to be able to perform simple matrix operations such as adding and multiplication
- to be able to find the reverse of the 2-dimensional matrix

7.1 Matrix

A matrix is a rectangular array of numbers e.g.

$$\mathbf{A} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & ... & x_{1n} \\ x_{21} & x_{22} & x_{23} & ... & x_{2n} \\ ... & ... & ... & ... & ... \\ x_{m1} & x_{m2} & x_{1m3} & ... & x_{mn} \end{bmatrix}$$

where:

- the notional subscripts in the typical element x_{ij} refers to its row and column location in the array, e.g. x_{12} refers to element in the first row and second column
- we say that matrix has m rows and n columns and the **dimension** of a matrix is defined as $m \times n$
- a matrix can be viewed as a set of column vectors or a set of row vectors
- a vector can be viewed as a matrix with only one column or with only one row

7.2 Special matrices

- A matrix with the same number of rows as columns, $m = n$, is said to be a **square matrix**
- A matrix that is not squared, $m \neq n$ is called **rectangular matrix**
- A **null matrix** is composed of all 0
- An **identity matrix**, denoted as $\mathbf{I}$ or $\mathbf{I_n}$, is a square matrix with 1's on the main diagonal and all other elements equal to 0, e.g. a three-dimensional identity matrix is

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- A square matrix is said to be **symmetric** if $x_{ij} = x_{ji}$ e.g.

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 2 \\ 4 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}$$

- A **diagonal matrix** is a square matrix whose non-diagonal entries are all zero, that is $x_{ij} = 0$ for $i \neq j$, e.g.

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

- An **upper-triangular matrix** is a square matrix in which all entries below the diagonal are 0, that is $x_{ij} = 0$ for $i < j$ e.g.

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 2 & 5 \\ 0 & 0 & 3 \end{bmatrix}$$

- A **lower-triangular matrix** is a square matrix in which all entries above the digonal are 0, that is hat is $x_{ij} = 0$ for $i > j$ e.g.

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

7.3 Matrix operations

- matrix $\mathbf{A} = \mathbf{B}$ if both matrices have exactly the same dimension and if each element of $\mathbf{A}$ equals to the corresponding element of e.g. $\mathbf{A} = \mathbf{B}$ if $\mathbf{A} = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 2 & 5 \\ 0 & 0 & 3 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 2 & 5 \\ 0 & 0 & 3 \end{bmatrix}$

- for any matrix $\mathbf{A}$ the **transpose**, denoted by $\mathbf{A}^\top$ or $\mathbf{A}'$, is obtained by interchanging rows and columns, e.g. given matrix $\mathbf{A} = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 2 & 5 \\ 0 & 0 & 3 \end{bmatrix}$ we have $\mathbf{A}^\top = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 2 & 0 \\ 4 & 5 & 3 \end{bmatrix}$. The transpose of a transpose of a matrix yield the original matrix, $\left(\mathbf{A}^\top\right)^\top = \mathbf{A}$
- we can **add** two matrices if they have the same dimension, e.g.

$$\mathbf{A} + \mathbf{B} = \mathbf{A} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} + \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} x_{11} + y_{11} & x_{12} + y_{12} \\ x_{21} + y_{21} & x_{22} + y_{22} \end{bmatrix}$$

- we can **multiply** a matrix by **a scalar** δ e.g.

$$\delta \cdot \mathbf{A} = \begin{bmatrix} \delta \cdot x_{11} & \delta \cdot x_{12} \\ \delta \cdot x_{21} & \delta \cdot x_{22} \end{bmatrix}$$

- we can **multiply two matrices** if they are **conformable**, i.e. first matrix has the same number of columns as the number of rows in the second matrix. We then can write:

$$\mathbf{C} = \mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix} + \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ y_{31} & y_{32} \end{bmatrix} = \begin{bmatrix} x_{11} \cdot y_{11} + x_{12} \cdot y_{21} + x_{13} \cdot y_{31} & x_{11} \cdot y_{12} + x_{12} \cdot y_{22} + x_{13} \cdot y_{32} \\ x_{21} \cdot y_{22} + x_{23} \cdot y_{21} + x_{13} \cdot y_{31} & x_{21} \cdot y_{12} + x_{23} \cdot y_{22} + x_{13} \cdot y_{32} \end{bmatrix}$$

7.4 Inverse of a matrix

For a square matrix $\mathbf{A}$ there may exist a matrix $\mathbf{B}$ such that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}$. An **inverse**, if it exists, is denoted as $\mathbf{A}^{-1}$ and we can rewrite the definition as

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}$$

where $\mathbf{I}$ is an identify matrix (equivalent to 1). There is no division for matrices, instead we can use inverse to multiply the matrix by an inverse, similar to when instead of dividing the number a by b we multiply a by reciprocal of $b = \frac{1}{b}$

For a 2-dimensional matrix we can follow the below formula for obtaining the inverse

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}^{-1} = \frac{1}{x_{11} \cdot x_{22} - x_{12} \cdot x_{21}} \cdot \begin{bmatrix} x_{22} & -x_{12} \\ -x_{21} & x_{11} \end{bmatrix}$$

7.5 Orthogonal matrix

- A matrix $\mathbf{A}$ for which $\mathbf{A}^\top = \mathbf{A^{-1}}$ is true is said to be **orthogonal**

Exercise 7.1. Given matrices

$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

a) what is the dimension of matrix $\mathbf{A}$?
b) what is $\mathbf{A}^\top$?
c) which of the matrices is i) an identity matrix ii) a square matrix, iii) null matrix, iv) diagonal matrix, v) a triangular matrix,?
d) calculate $\mathbf{A} + \mathbf{B}$?
e) calculate $\mathbf{A} \cdot \mathbf{C}$?
f) calculate $\mathbf{B}^\top$
g) calculate $\mathbf{A}^{-1}$
h) calculate $(\mathbf{A} + \mathbf{B})^{-1}$

Answers to selected exercises (matrices)

Exr. 7.1

a) 2×2

b) $\mathbf{A}^\top = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$

c) i) identity matrix: $\mathbf{B}$, ii) a square matrix: $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, iii) null matrix: none, iv) diagonal matrix: $\mathbf{B}$ (identity matrix is diagonal) and $\mathbf{C}$, v) triangular $\mathbf{B}$ and $\mathbf{C}$ as both identify matrix $\mathbf{B}$ and diagonal matrix $\mathbf{C}$ is triangular, both lower and upper triangular

d) $\mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 3 & 5 \end{bmatrix}$

e) $\mathbf{A} \cdot \mathbf{C} = \begin{bmatrix} 1 \cdot 1 + 2 \cdot 0 & 1 \cdot 0 + 2 \cdot 2 \\ 3 \cdot 1 + 4 \cdot 0 & 3 \cdot 0 + 4 \cdot 2 \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 3 & 8 \end{bmatrix}$

f) $\mathbf{B}^\top = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

g)

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^{-1} = \frac{1}{1 \cdot 4 - 2 \cdot 3} \cdot \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} = -\frac{1}{2} \cdot \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}$$

Part II

Linear Models

Chapter 8

Introduction to linear models

Aims

- to introduce concept of linear models using simple linear regression

Learning outcomes

- to understand what a linear model is and be familiar with the terminology
- to be able to state linear model in the general vector-matrix notation
- to be able to use the general vector-matrix notation to numerically estimate model parameters
- to be able to use `lm()` function for model fitting, parameter estimation, hypothesis testing and prediction

8.1 Statistical vs. deterministic relationship

Relationships in probability and statistics can generally be one of three things: deterministic, random, or statistical:

- a **deterministic** relationship involves **an exact relationship** between two variables, for instance Fahrenheit and Celsius degrees is defined by an equation $Fahrenheit = \frac{9}{5} \cdot Celcius + 32$
- there is **no relationship** between variables in the **random relationship**, for instance number of succulents Olga buys and time of the year as Olga keeps buying succulents whenever she feels like it throughout the entire year
- **a statistical relationship** is a **mixture of deterministic and random relationship**, e.g. the savings that Olga has left in the bank account depend on Olga's monthly salary income (deterministic part) and the money

spent on buying succulents (random part)

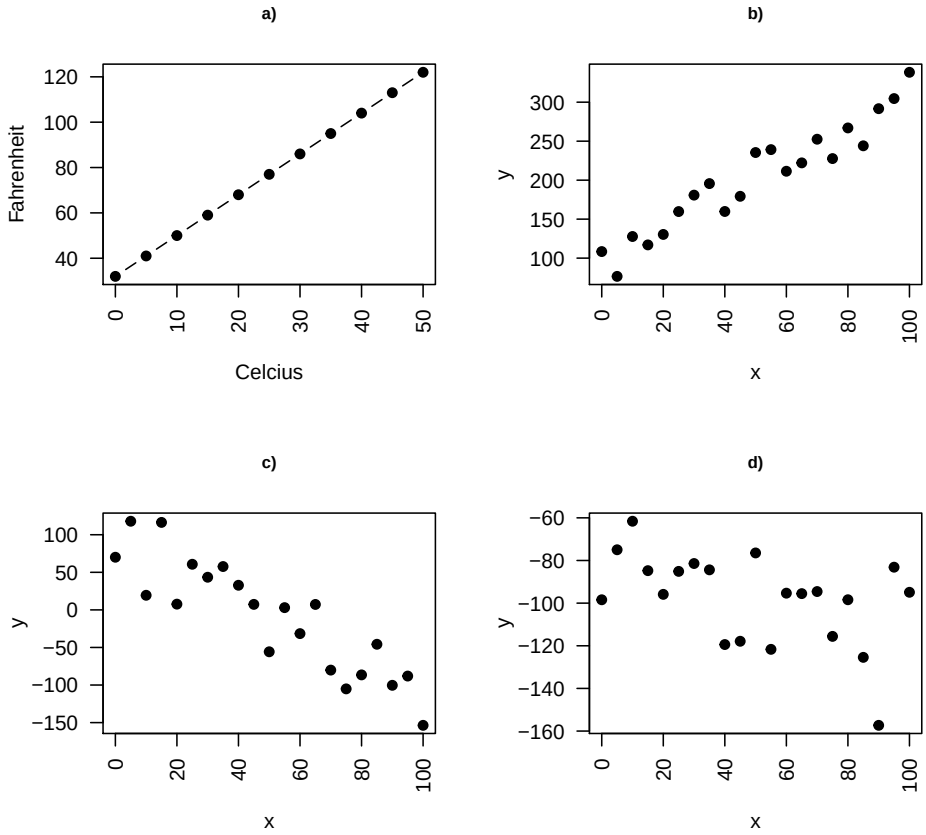


Figure 8.1: Deterministic vs. statistical relationship: a) deterministic: equation exactly describes the relationship between the two variables e.g. Ferenheit and Celcius relationship ; b) statistical relationship between x and y is not perfect (increasing relationship), c) statistical relationship between x and y is not perfect (decreasing relationship), d) random signal

8.2 What linear models are and are not

- A linear model is one in which the parameters appear linearly in the deterministic part of the model
- e.g. **simple linear regression** through the origin is a simple linear model of the form $Y_i = \beta x + \epsilon$ often used to express a relationship of one numerical variable to another, e.g. the calories burnt and the kilometers cycled
- linear models can become quite advanced by including more variables, e.g. the calories burnt could be a function of both the kilometers cycled and status of bike, or the transformation of the variables

More examples where model parameters appear linearly:

- $Y_i = \alpha + \beta x_i + \gamma x_i + \epsilon_i$
- $Y_i = \alpha + \beta x_i^2 \epsilon$
- $Y_i = \alpha + \beta x_i^2 + \gamma x_i^3 + \epsilon$

and an example on a non-linear model where parameter β appears in the exponent of x_i

- $Y_i = \alpha + x_i^{\beta} + \epsilon$

8.3 Terminology

There are many terms and notations used interchangeably:

- y is being called:
  - response
  - outcome
  - dependent variable
- x is being called:
  - exposure
  - explanatory variable
  - dependent variable
  - predictor
  - covariate

8.4 With linear models we can answer questions such as:

- is there a relationship between exposure and outcome, e.g. body weight and plasma volume?
- how strong is the relationship between the two variables?
- what will be a predicted value of the outcome given a new set of exposure values?
- how accurately can we predict outcome?
- which variables are associated with the response, e.g. is it body weight and height that can explain the plasma volume or is it just the body weight?

8.5 Simple linear regression

- It is used to estimate the best-fitting straight line to describe the association between the outcome and one explanatory variable
- For example, let's look at the example data containing body weight (kg) and plasma volume (liters) for eight healthy men to see what it the best-fitting straight line is.

Example data:

```
weight <- c(58, 70, 74, 63.5, 62.0, 70.5, 71.0, 66.0) # body weight (kg)
plasma <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12) # plasma volume (liters)
```

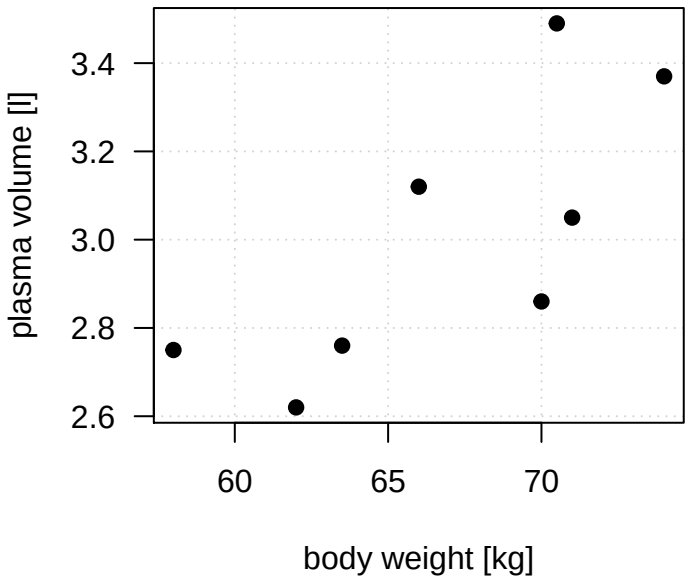


Figure 8.2: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice verca*.

The equation for the red line is:

$$Y_i = 0.086 + 0.044 \cdot x_i \quad for\ i = 1 \ldots 8$$

and in general:

$$Y_i = \alpha + \beta \cdot x_i \quad for\ i = 1 \ldots n$$

- In other words, by finding the best-fitting straight line we are **building a statistical model** to represent the relationship between plasma volume (Y) and explanatory body weight variable (x)
- If were to use our model $Y_i = 0.086 + 0.044 \cdot x_i$ to find plasma volume given a weight of 58 kg (our first observation, $i = 1$), we would notice that we would get $Y = 0.086 + 0.044 \cdot 58 = 2.638$, not exactly 2.75 as we have for our first man in our dataset that we started with, i.e. $2.75 - 2.638 = 0.112 \neq 0$.
- We thus add to the above equation an **error term** to account for this and now we can write our **simple regression model** more formally as:

$$Y_i = \alpha + \beta \cdot x_i + \epsilon_i \tag{8.1}$$

- where we call α and β **model coefficients**
- where we call ϵ_i **error terms**

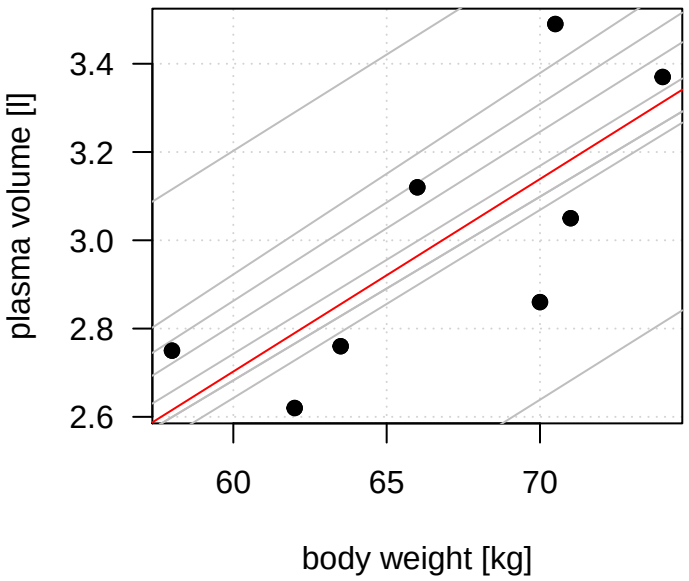


Figure 8.3: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice verca*. Linear regression gives the equation of the straight line (red) that best describes how the outcome changes (increase or decreases) with a change of exposure variable

8.6 Least squares

- in the above body weight - plasma volume example, the values of α and β have just appeared
- in practice, α and β values are unknown and we use data to estimate these coefficients, noting the estimates with a hat, $\hat{\alpha}$ and $\hat{\beta}$
- **least squares** is one of the methods of parameters estimation, i.e. finding $\hat{\alpha}$ and $\hat{\beta}$

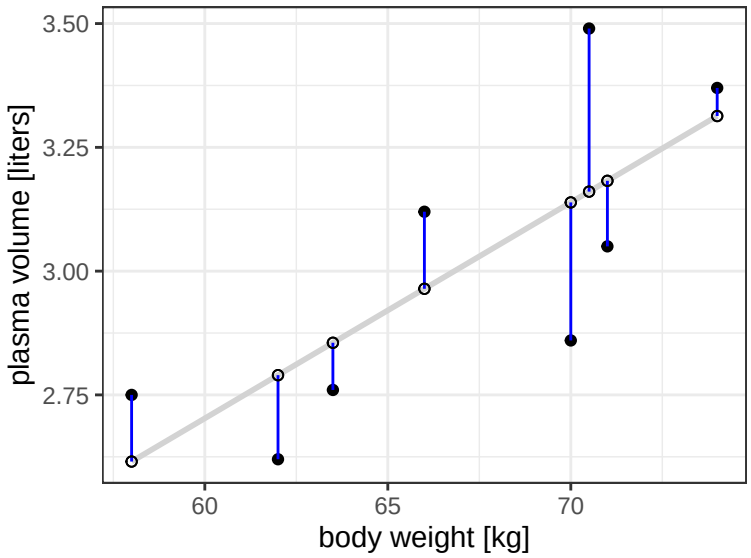


Figure 8.4: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice verca*. Linear regrssion gives the equation of the straight line (red) that best describes how the outcome changes with a change of exposure variable. Blue lines represent error terms, the vertical distances to the regression line

Let $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ be the prediction y_i based on the i-th value of x:

- Then $\epsilon_i = y_i - \hat{y}_i$ represents the i-th **residual**, i.e. the difference between the i-th observed response value and the i-th response value that is predicted by the linear model
- RSS, the **residual sum of squares** is defined as:

$$RSS = \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_n^2$$

or equivalently as:

$$RSS = (y_1 - \hat{\alpha} - \hat{\beta}x_1)^2 + (y_2 - \hat{\alpha} - \hat{\beta}x_2)^2 + ... + (y_n - \hat{\alpha} - \hat{\beta}x_n)^2$$

- the least squares approach chooses $\hat{\alpha}$ and $\hat{\beta}$ **to minimize the RSS**. With some calculus we get Theorem 8.1

Theorem 8.1 (Least squares estimates for a simple linear regression)**.**

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \frac{S_{xy}}{S_{xx}} \cdot \bar{x}$$

where:

- $\bar{x}$*: mean value of* x
- $\bar{y}$*: mean value of* y
- Sxx*: sum of squares of* X *defined as* $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$
- Syy*: sum of squares of* Y *defined as* $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Sxy*: sum of products of* X *and* Y *defined as* $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

We can further re-write the above sum of squares to obtain

- sum of squares of X,

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

- sum of products of X and Y

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

Example (Least squares)

Let's try least squares method to find coefficient estimates in our body weight and plasma volume example

```
# initial data
weight <- c(58, 70, 74, 63.5, 62.0, 70.5, 71.0, 66.0) # body weight (kg)
plasma <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12) # plasma volume (liters)

# rename variables for convenience
x <- weight
```

```
y <- plasma

# mean values of x and y
x.bar <- mean(x)
y.bar <- mean(y)

# Sum of squares
Sxx <-  sum((x - x.bar)^2)
Sxy <- sum((x-x.bar)*(y-y.bar))

# Coefficient estimates
beta.hat <- Sxy / Sxx
alpha.hat <- y.bar - Sxy/Sxx*x.bar

# Print estimated coefficients alpha and beta
print(alpha.hat)
## [1] 0.08572428

print(beta.hat)
## [1] 0.04361534
```

In R we can use `lm`, the built-in function, to fit a linear regression model and we can replace the above code with one line

```
lm(plasma ~ weight)
##
## Call:
## lm(formula = plasma ~ weight)
##
## Coefficients:
## (Intercept)       weight
##     0.08572      0.04362
```

8.7 Intercept and Slope

- Linear regression gives us estimates of model coefficient $Y_i = \alpha + \beta x_i + \epsilon_i$
- α is known as the intercept
- β is known as the slope

8.8 Hypothesis testing

- the calculated $\hat{\alpha}$ and $\hat{\beta}$ are estimates of the population values of the intercept and slope and are, therefore, subject to sampling variation
- their precision is measure by their standard errors, e.s.e($\hat{\alpha}$) and e.s.e($\hat{\beta}$)

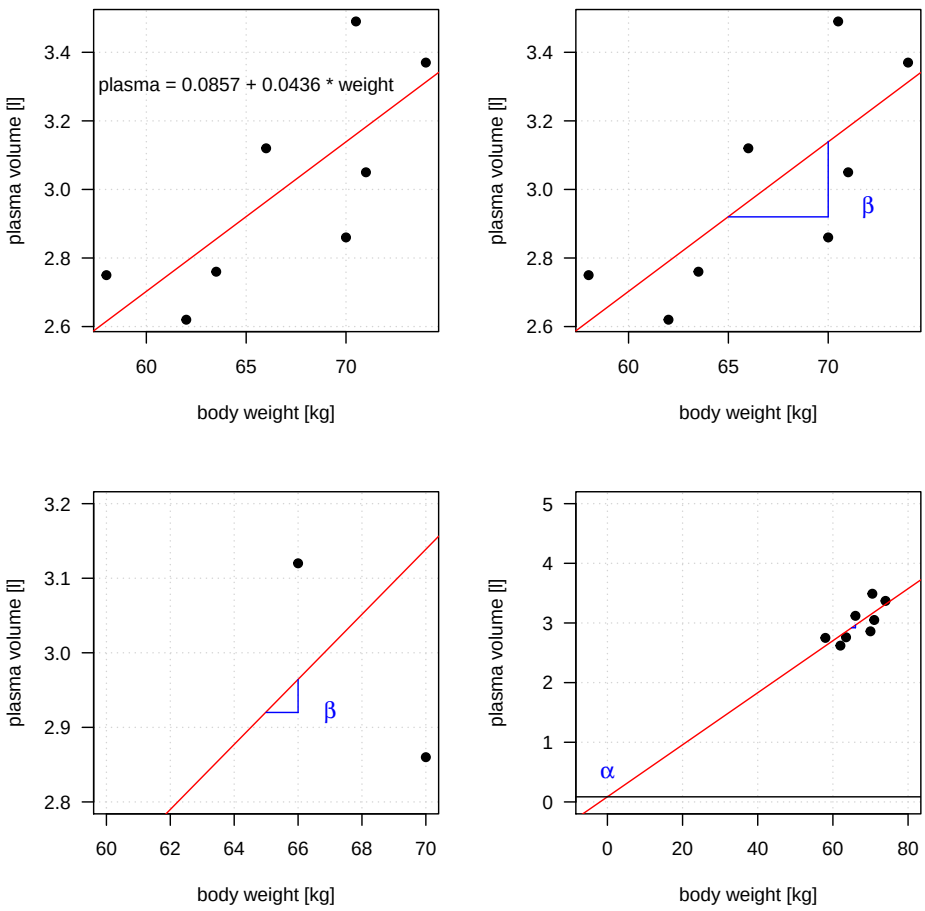


Figure 8.5: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice verca*. Linear regression gives the equation of the straight line that best describes how the outcome changes (increase or decreases) with a change of exposure variable (in red)
(#fig:lm-parameters, fig-intro-example-reg-parameters)

- the estimated standard errors are used in hypothesis testing, confidence and prediction intervals

The most common hypothesis test involves testing the `null hypothesis` of:

- H_0 : There is no relationship between X and Y
- versus the `alternative hypothesis` H_a : there is some relationship between X and Y

Mathematically, this corresponds to testing:

- $H_0 : \beta = 0$
- versus $H_0 : \beta \neq 0$
- since if $\beta = 0$ then the model $Y_i = \alpha + \beta x_i + \epsilon_i$ reduces to $Y = \alpha + \epsilon_i$

Under the null hypothesis:

- $H_0 : \beta = 0$ we have: $\frac{\hat{\beta} - \beta}{e.s.e(\hat{\beta})} \sim t(n-p)$, where
- n is number of observations
- p is number of model parameters
- $\frac{\hat{\beta} - \beta}{e.s.e(\hat{\beta})}$ is called the t-statistics
- that follows Student's t distribution with $n - p$ degrees of freedom

Example (Hypothesis testing)

Let's look again at our example data. This time we will not only fit the linear regression model but look a bit more closely at the R summary of the model

```
weight <- c(58, 70, 74, 63.5, 62.0, 70.5, 71.0, 66.0) # body weight (kg)
plasma <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12) # plasma volume (liters)

model <- lm(plasma ~ weight)
print(summary(model))
##
## Call:
## lm(formula = plasma ~ weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27880 -0.14178 -0.01928  0.13986  0.32939
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08572    1.02400   0.084   0.9360
## weight       0.04362    0.01527   2.857   0.0289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2188 on 6 degrees of freedom
```

```
## Multiple R-squared:  0.5763,         Adjusted R-squared:  0.5057
## F-statistic:  8.16 on 1 and 6 DF,  p-value: 0.02893
```

- Under "Estimate" we see estimates of our model coefficients, $\hat{\alpha}$ (intercept) and $\hat{\beta}$ (here weight), followed by their estimated standard errors.
- If we were to test if there is an association between weight and plasma volume we would write under $H_0 : \beta = 0$ and $\frac{\hat{\beta}-\beta}{e.s.e(\hat{\beta})} = \frac{0.04362-0}{0.01527} = 2.856582$
- and we would compare t-statistics to Student's t distribution with $n-p = 8-2 = 6$ degrees of freedom (we have two model parameters, α and β)
- we can use Student's t distribution table or R code to obtain p-value

```
2*pt(2.856582, df=6, lower=F)
## [1] 0.02893095
```

- here the observed t-statistics is large and therefore yields a small p-value, meaning that there is sufficient evidence to reject null hypothesis in favor of the alternative and conclude that there is an significant association between weight and plasma volume

8.9 Vector-matrix notations

While in simple linear regression it is feasible to arrive at the parameters estimates using calculus in more realistic settings with multiple regression (more than one explanatory variable in the model) it is more efficient to use vectors and matrices to define the regression model.

Let's rewrite our simple linear regression model $Y_i = \alpha + \beta_i + \epsilon_i \quad i = 1, \dots n$ into vector-matrix notations.

- First we rename our α to β_0 and β to β_1 (it is easier to keep tracking the number of model parameters this way)
- Then we notice that we actually have n equations such as:

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + \epsilon_2$$

$$y_3 = \beta_0 + \beta_1 x_3 + \epsilon_3$$

$$\dots$$

$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n$$

- we can group all Y_i and ϵ_i into column vectors: $\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ and $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

- we stack two parameters β_0 and β_1 into another column vector:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

- we then append a vector of ones with the single predictor for each i and create a matrix with two columns: design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Now we can write our linear model in a vector-matrix notations as:

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon$$

Definition: vector matrix form of the linear model

The vector-matrix representation of a linear model with $p-1$ predictors can be written as

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon$$

where:

- $\mathbf{Y}$ is $n \times 1$ vector of observations
- β is $p \times 1$ vector of parameters
- $\mathbf{X}$ is $n \times p$ design matrix
- ϵ is $n \times 1$ vector of vector of random errors, indepedent and identically distributed (i.i.d) N(0, σ^2)

In full, the above vectors and matrix have the form:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & \dots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix}$$

Theorem 8.2 (Least squares in vector-matrix notation)**.** *The least squares estimates for a linear regression of the form:*

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon$$

is given by:

$$\hat{} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Example: vector-matrix notation

Following the above definition we can write our weight - plasma volume model as:

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon$$

where:

$$\mathbf{Y} = \begin{bmatrix} 2.75 \\ 2.86 \\ 3.37 \\ 2.76 \\ 2.62 \\ 3.49 \\ 3.05 \\ 3.12 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_8 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 58.0 \\ 1 & 70.0 \\ 1 & 74.0 \\ 1 & 63.5 \\ 1 & 62.0 \\ 1 & 70.5 \\ 1 & 71.0 \\ 1 & 66.0 \end{bmatrix}$$

and we can estimate model parameters using $\hat{} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ in R (although we could try solving it by hand)

```
n <- length(plasma) # no. of observation
Y <- as.matrix(plasma, ncol=1)
X <- cbind(rep(1, length=n), weight)
X <- as.matrix(X)

# print Y and X to double-check that the format is according to the definition
print(Y)
##      [,1]
## [1,] 2.75
## [2,] 2.86
## [3,] 3.37
## [4,] 2.76
## [5,] 2.62
## [6,] 3.49
## [7,] 3.05
## [8,] 3.12
print(X)
##        weight
## [1,] 1   58.0
## [2,] 1   70.0
```

```
## [3,] 1   74.0
## [4,] 1   63.5
## [5,] 1   62.0
## [6,] 1   70.5
## [7,] 1   71.0
## [8,] 1   66.0

# least squares estimate
beta.hat <- solve(t(X)%*%X)%*%t(X)%*%Y # solve() finds inverse of matrix
print(beta.hat)
##               [,1]
##        0.08572428
## weight 0.04361534
```

8.10 Confidence intervals and prediction intervals

- when we estimate coefficients we can also find their **confidence intervals**, e.g. 95% confidence intervals, i.e. a range of vales that contain the true unknown value of the parameter
- we we use linear regression models to predict the response value given a new observation, we can find **prediction intervals**. Here, we look at any specific value of x_i, and find an interval around the predicted value y_i' for x_i such that there is a 95% probability that the real value of y (in the population) corresponding to x_i is within this interval

Before we said that we use estimated standard error in hypothesis testing and finding the intervals but we have not yet said how to calculate e.s.e. Using vector-matrix notation we can now write that:

$$\frac{(\mathbf{b}\hat{\beta} - \mathbf{b}^T\beta)}{\sqrt{\frac{RSS}{n-p}\mathbf{b^T}(\mathbf{X^T X})^{-1}\mathbf{b}}}$$

where:

- the denominator would yield e.s.e(β_1) if $\mathbf{b^T} = (0 \quad 1)$ and a model $Y_i = \beta_0 + \beta_1 x + \epsilon_i$
- a confidence interval estimate for β_1 could be estimated via:

$$\mathbf{b^T}\hat{\beta} \pm (n-p; \frac{1+c}{2}\sqrt{\frac{RSS}{n-p}}(\mathbf{b^T}(\mathbf{X^T X})^{-1}\mathbf{b}))$$

- and a prediction interval with confidence c is

$$\mathbf{b^T}\hat{\beta} \pm (n-p; \frac{1+c}{2}\sqrt{\frac{RSS}{n-p}}(1 + \mathbf{b^T}(\mathbf{X^T X})^{-1}\mathbf{b})$$

We will not go further into these calculations here:

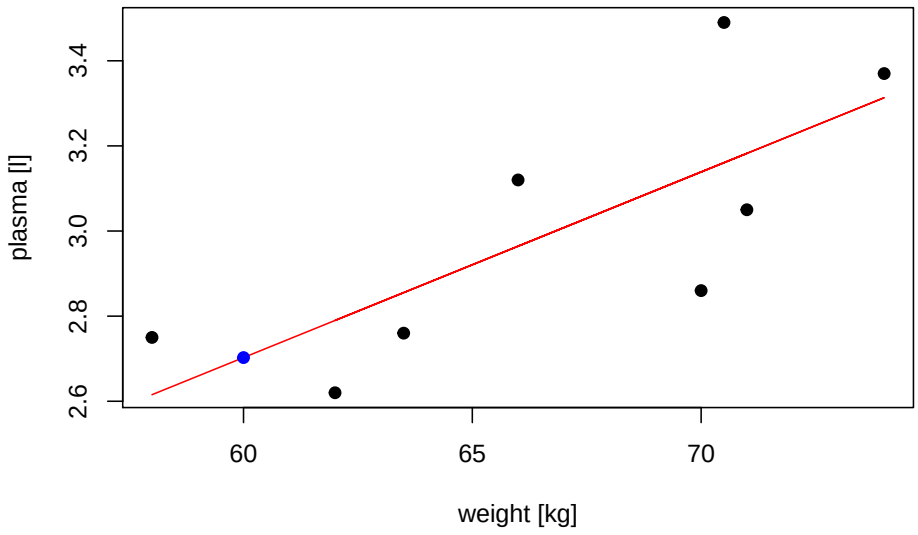
- just remember that the prediction interval > than a confidence interval
- note (1 +) in the prediction interval equation

Example: prediction and intervals

Let's: - find confidence intervals for our coefficient estimates - predict plasma volume for a men weighting 60 kg - find prediction interval - plot original data, fitted regression model, predicted observation

```
# fit regression model
model <- lm(plasma ~ weight)
print(summary(model))
##
## Call:
## lm(formula = plasma ~ weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27880 -0.14178 -0.01928  0.13986  0.32939
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08572    1.02400   0.084   0.9360
## weight       0.04362    0.01527   2.857   0.0289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2188 on 6 degrees of freedom
## Multiple R-squared:  0.5763,        Adjusted R-squared:  0.5057
## F-statistic:  8.16 on 1 and 6 DF,  p-value: 0.02893

# find confidence intervals for the model coefficients
confint(model)
##                    2.5 %     97.5 %
## (Intercept) -2.419908594 2.59135716
## weight       0.006255005 0.08097567

# predict plasma volume for a new observation of 60 kg
# we have to create data frame with a variable name matching the one used to build the model
new.obs <- data.frame(weight = 60)
```

```r
predict(model, newdata = new.obs)
##        1
## 2.702645

# find prediction intervals
predict(model, newdata = new.obs,  interval = "prediction")
##        fit      lwr      upr
## 1 2.702645 2.079373 3.325916

# plot the original data, fitted regression and predicted value
plot(weight, plasma, pch=19, xlab="weight [kg]", ylab="plasma [l]")
lines(weight, model$fitted.values, col="red") # fitted model in red
points(new.obs, predict(model, newdata = new.obs), pch=19, col="blue") # predicted val
```



8.11 Exercises: linear models I

Exercise 8.1. Linear models form

Which of the following models are linear models and why?

a) $Y_i = \alpha + \beta x_i + \epsilon_i$
b) $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$
c) $Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i$
d) $Y_i = \alpha + \gamma x_i^\beta + \epsilon_i$

Exercise 8.2. Protein levels in pregnancy

The researchers were interested whether protein levels in expectant mothers are changing throughout the pregnancy. Observations have been taken on 19 healthy women and each woman was at different stage of pregnancy (gestation).

Assuming linear model:

- $Y_i = \alpha + \beta x_i + \epsilon_i$, where Y_i corresponds to protein levels in i-th observation

and taking summary statisitcs:

- $\sum_{i=1}^{n} x_i = 456$
- $\sum_{i=1}^{n} x_i^2 = 12164$
- $\sum_{i=1}^{n} x_i y_i = 369.87$
- $\sum_{i=1}^{n} y_i^2 = 11.55$

a) find the least square estimates of $\hat{\alpha}$ and $\hat{\beta}$
b) knowing that e.s.e$(\hat{\beta}) = 0.022844$ can we

- i) reject the null hypothesis that the is no relationship between protein level and gestation, i.e. perform a hypothesis test to test $H_0 : \beta = 0$;

- ii) can we reject the null hypothesis that $\beta = 0.02$, i.e. perform a hypothesis test to test $H_0 : \beta = 0.02$

c) write down the linear model in the vector-matrix notation and identify response, parameter, design and error matrices
d) read in "protein.csv" data into R, set Y as protein (response) and calculate using matrix functions the least squares estimates of model coefficients
e) use `lm()` function in R to check your calculations
f) use the fitted model in R to predict the value of protein levels at week 20. Try plotting the data, fitted linear model and the predicted value to assess whether your prediction is to be expected.

Exercise 8.3. The glucose level in potatoes depends on their storage time and the relationship is somehow curvilinear as shown below. As we believe that the quadratic function might describe the relationship, assume linear model in form $Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i \quad i = 1, \ldots, n$ where $n = 14$ and

a) write down the model in vector-matrix notation
b) load data to from "potatoes.csv" and use least squares estimates for obtain estimates of model coefficients
c) perform a hypothesis test to test $H_0 : \gamma = 0$; and comment whether we there is a significant quadratic term
d) use `lm()` function to verify your calculations
e) predict glucose concentration at storage time 4 and 16 weeks. Plot the data, fitted model and predicted values

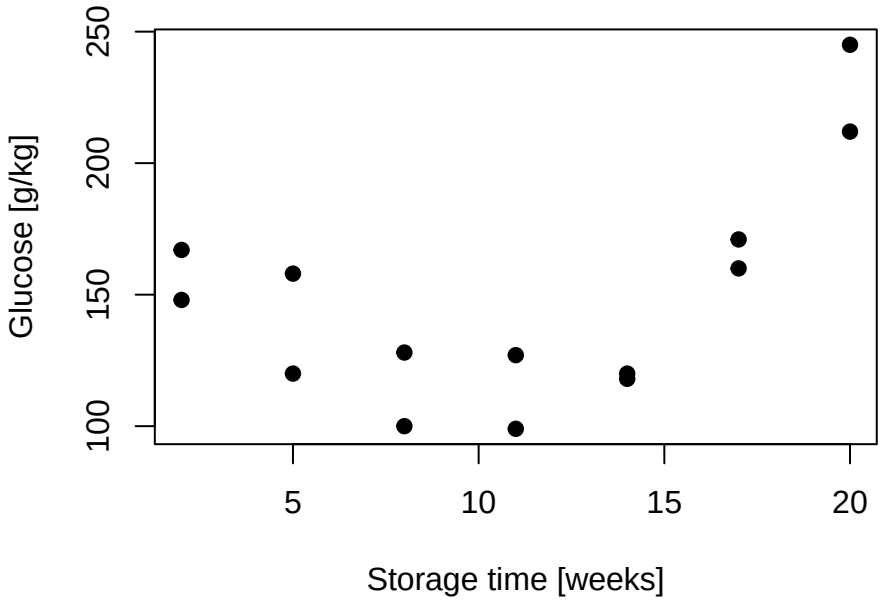


Figure 8.6: Sugar in potatoes: relationship between storage time and glucose content

Answers to selected exercises (linear models)

Exr. 8.2

a)

- $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}) = 12164 - \frac{456^2}{19} = 1220$
- $S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 369.87 - \frac{(456 \cdot 14.25)}{19} = 27.87$
- $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = 27.87/1220 = 0.02284$
- $\hat{\alpha} = \bar{y} - \frac{S_{xy}}{S_{xx}} \cdot \bar{x} = \frac{14.25}{19} - \frac{27.87}{1220} \cdot \frac{456}{19} = 0.20174$

b) i.

We can calculate test-statistics following:

- $\frac{\hat{\beta}-\beta}{e.s.e(\hat{\beta})} \sim t(n-p) = \frac{0.02284-0}{0.20174} = 6.934$ where the value follows Student's t distribution with $n - p = 19 - 2 = 17$ degrees of freedom. We can now estimate the a p-value using Student's t distribution table or use a function in R

```
2*pt(6.934, df=17, lower=F)
## [1] 2.414315e-06
```

As p-value « 0.001 there is sufficient evidence to reject H_0 in favor of H_1, thus

we can conclude that there is a significant relationship between protein levels and gestation

b) ii.

Similarly, we can test $H_0 : \beta = 0.02$, i.e. $\frac{\hat{\beta}-\beta}{e.s.e(\hat{\beta})} \sim t(n-p) = \frac{0.02284-0.02}{0.20174} = 0.01407753$. Now the test statistics is small

```
2*pt(0.01407753, df=17, lower=F)
## [1] 0.988932
```

p-value is large and hence there is no sufficient evidence to reject H_0 and we can conclude that $\beta = 0.02$

c) We can rewrite the linear model in vector-matrix formation as $\mathbf{Y} = \mathbf{X} +$ where:

$$\text{response } \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{19} \end{bmatrix}$$

$$\text{parameters } = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

$$\text{design matrix } \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{19} \end{bmatrix}$$

$$\text{errors } = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{19} \end{bmatrix}$$

d) The least squares estimates in vector-matrix notation is $\hat{} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and we can calculate this in R

```
# read in data
data.protein <- read.csv("data/lm/protein.csv")

# print out top observations
head(data.protein)
##   Protein Gestation
## 1    0.38        11
## 2    0.58        12
## 3    0.51        13
## 4    0.38        15
## 5    0.58        17
## 6    0.67        18
```

```r
# define Y and X matrices given the data
n <- nrow(data.protein) # nu. of observations
Y <-  as.matrix(data.protein$Protein, ncol=1) # response
X <-  as.matrix(cbind(rep(1, length=n), data.protein$Gestation)) # design matrix
head(X) # double check that the design matrix looks like it should
##      [,1] [,2]
## [1,]    1   11
## [2,]    1   12
## [3,]    1   13
## [4,]    1   15
## [5,]    1   17
## [6,]    1   18

# least squares estimate
beta.hat <- solve(t(X)%*%X)%*%t(X)%*%Y
print(beta.hat)
##             [,1]
## [1,] 0.20173770
## [2,] 0.02284426
```

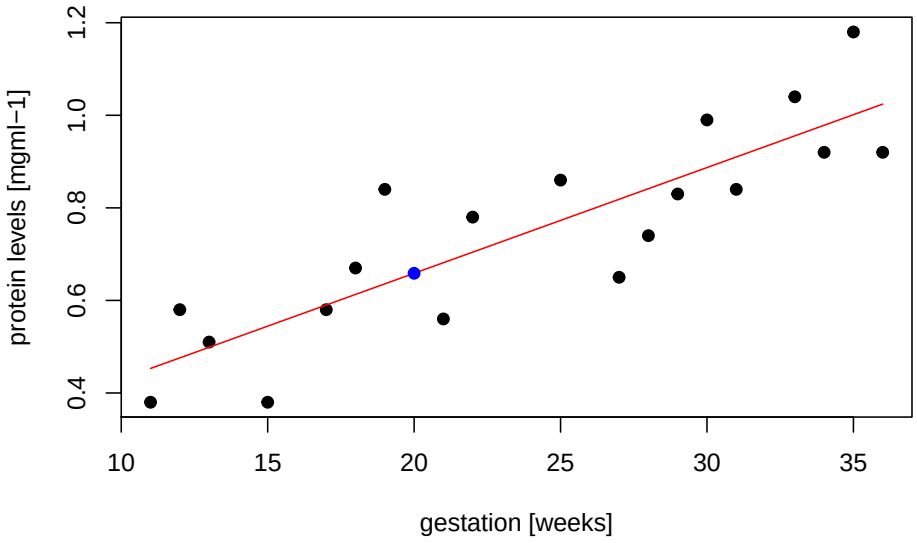
e) We use `lm()` function to check our calculations

```r
# fit linear regression model and print model summary
protein <- data.protein$Protein # our Y
gestation <- data.protein$Gestation # our X

model <- lm(protein ~ gestation)
print(summary(model))
##
## Call:
## lm(formula = protein ~ gestation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16853 -0.08720 -0.01009  0.08578  0.20422
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.201738   0.083363   2.420    0.027 *
## gestation   0.022844   0.003295   6.934 2.42e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1151 on 17 degrees of freedom
## Multiple R-squared:  0.7388,        Adjusted R-squared:  0.7234
```

```
## F-statistic: 48.08 on 1 and 17 DF,  p-value: 2.416e-06
```

f)

```
new.obs <- data.frame(gestation = 20)
y.pred <- predict(model, newdata = new.obs)

# we can visualize the data, fitted linear model (red), and the predicted value (blue)
plot(gestation, protein, pch=19, xlab="gestation [weeks]", ylab="protein levels [mgml-1]")
lines(gestation, model$fitted.values, col="red")
points(new.obs, y.pred, col="blue", pch=19, cex = 1)
```



Exr. 8.3

a) We can rewrite the linear model in vector-matrix formation as $\mathbf{Y} = \mathbf{X} +$ where:

$$\text{response } \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{14} \end{bmatrix}$$

$$\text{parameters } = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}$$

$$\text{design matrix } \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{14} & x_{14}^2 \end{bmatrix}$$

$$\text{errors} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{14} \end{bmatrix}$$

b) load data to from "potatoes.csv" and use least squares estimates for obtain estimates of model coefficients

```
data.potatoes <- read.csv("data/lm/potatoes.csv")

# define matrices
n <- nrow(data.potatoes)
Y <-  data.potatoes$Glucose
X1 <- data.potatoes$Weeks
X2 <- (data.potatoes$Weeks)^2
X <- cbind(rep(1, length(n)), X1, X2)
X <- as.matrix(X)

# least squares estimate
# beta here refers to the matrix of model coefficients incl. alpha, beta and gamma
beta.hat <- solve(t(X)%*%X)%*%t(X)%*%Y
print(beta.hat)
##          [,1]
##    200.169312
## X1 -19.443122
## X2   1.030423
```

c) we use `lm()` function to verify our calculations:

```
model <- lm(Y ~ X1 + X2)
print(summary(model))
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.405 -11.250  -8.071  12.911  29.286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 200.1693    15.0527  13.298 4.02e-08 ***
## X1          -19.4431     3.1780  -6.118 7.54e-05 ***
## X2            1.0304     0.1406   7.329 1.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 16.4 on 11 degrees of freedom
## Multiple R-squared:  0.8694,Adjusted R-squared:  0.8457
## F-statistic: 36.61 on 2 and 11 DF,  p-value: 1.373e-05
```

d) perform a hypothesis test to test $H_0 : \gamma = 0$; and comment whether we there is a significant quadratic term

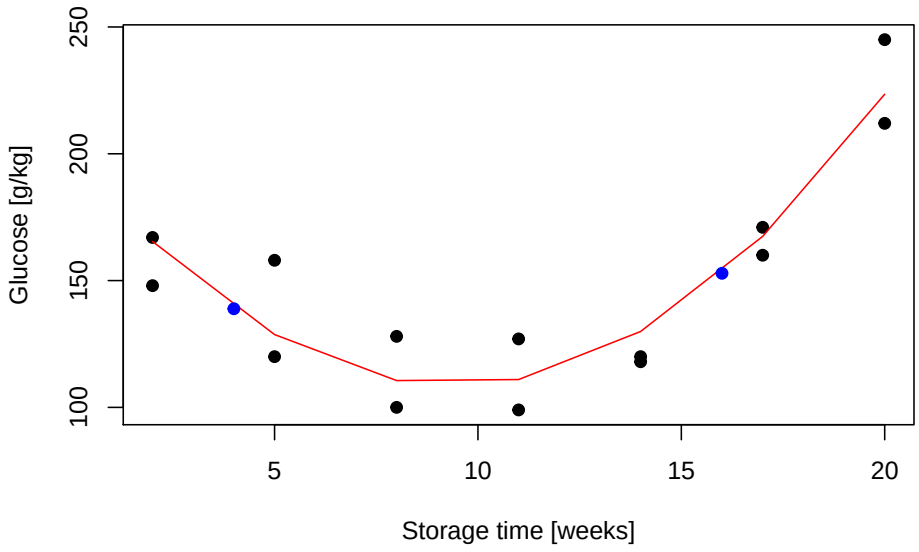
- $\frac{\hat{\gamma}-\gamma}{e.s.e(\hat{\gamma})} \sim t(n-p) = \frac{1.030423-0}{0.1406} = 7.328755$ where the value follows Student's t distribution with $n-p = 19-2 = 17$ degrees of freedom. We can now estimate the a p-value using Student's t distribution table or use a function in R

```
2*pt(7.328755, df=14-3, lower=F)
## [1] 1.487682e-05
```

As p-value « 0.001 there is sufficient evidence to reject H_0 in favor of H_1, thus we can conclude that there is a significant quadratic relationship between glucose and storage time

e) predict glucose concentration at storage time 4 and 16 weeks

```
new.obs <- data.frame(X1 = c(4, 16), X2 = c(4^2, 16^2))
pred.y <- predict(model, newdata = new.obs)

plot(data.potatoes$Weeks, data.potatoes$Glucose, xlab="Storage time [weeks]", ylab="Glucose [g/kg
lines(data.potatoes$Weeks, model$fitted.values, col="red")
points(new.obs[,1], pred.y, pch=19, col="blue")
```



Chapter 9

Interpreting regression coefficients

Aims

- xxx

Learning outcomes

- xxx

Chapter 10

Model assumptions

Aims

- xxx

Learning outcomes

- xxx

Chapter 11

Generalized linear models

Aims

- xxx

Learning outcomes

- xxx

Chapter 12

Linear Mixed Models

Aims

- xxx

Learning outcomes

- xxx