INTRODUCTION TO MODERN MATHEMATICAL MODELING

INTRODUCTION TO MODERN MATHEMATICAL MODELING

-Developing Mathematical Models Using Data

Samuel S.P. Shen
San Diego State University

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

CONTENTS

FOREWORD

This is primarily a graduate mathematical modeling textbook, but can also be used as a handbook for research in mathematics and statistics applications to natural sciences, engineering, social sciences, and applied mathematics itself. It includes basic skills of applied mathematics consulting, such as dimensional analysis for exploring possible relationships among the relevant variables, R programs for basic statistics, linear algebra, and space-time plotting, and 5-step PAMMI () method of mathematical modeling principles.

The prerequisite for this course are Calculus I-III, the first semester of linear algebra, and some initial background of ODE (ordinary differential equations), complex variables, and Fourier series. Thus, students are assumed to know gradient, divergence theorem, vector calculus, spherical coordinates, Cauchy-Riemann condition for analytic functions of a complex variable, SVD(singular value decomposition) of a rectangular matrix, eigenvalues and eigenvectors of square matrices, dot products, and cross product. However, this text still reviews these concepts, definitions, critical formulas and main theorems of the above topics when being used, since some students are rusty with the calculations and have never not fully understood the meaning o the mathematical results anyway. ========= Topics covered in this book includes linear regression models, linear algebra models, probability models, calculus models, differential equation models, stochastic models, machine learning models, big data models, dimensional analysis, and R programs.

R programming is taught in class from beginning, although other computer program languages are allowed. R is free for public download and can be installed easily for either PC or Mac. Computer programming experience is not required for reading this book.

Mathematical model is a mathematical expression, often a formula or an equation, that describes a phenomenon, such as the free-fall of an object from a height. The distance between of the object and its initial release position is modeled by $(1/2)gt^2$, where g is the gravitational acceleration and t is the time from the release. Science history implies that Galileo Galilei (1564-1642) was the first who invented this formula. He designed a very smart experiment for this. At that time, it was hard to observe the free time since a body falls down too fast in the free fall environment. He slowed down the fall by free roll of a ball on a plate with ticks (see Fig. 0.1). He placed a wire on the plate so that the ball would make a click sound when the ball rolled over the wire. He adjusted the positions of the four wires so that the ball would make click sound in uniform time intervals. He then discovered that the distance after each click sound is

$$(1/2)at^2\ [m] \tag{0.1}$$

where $a = g \sin\theta$ and θ is the angle between the plate and the horizontal plane. The four lines' distances from the releasing points are thus

$$0.5a \times 1^2, \quad 0.5a \times 2^2, \quad 0.5a \times 3^2, \quad 0.5a \times 4^2\ [m]. \tag{0.2}$$

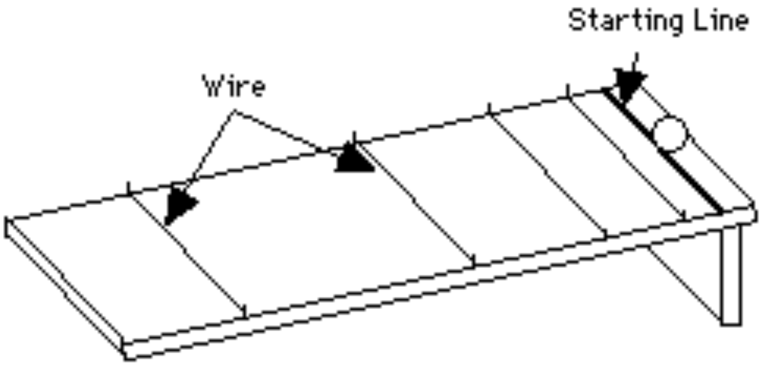


Figure 0.1 Galileo's experiment for a ball falling down on an inclined plate.

The formula $s = (1/2)at^2$ is a mathematical model for the ball rolling down on a plate under gravity. Because of measurement errors, the model is not 100% accurate when compared with the obtained data of time and distance. The real world problem is often that when a certain phenomenon is observed, a mathematical model is needed to describe the phenomenon in a quantitative fashion, as accurately as one can. Because observations are necessarily involved most natural phenomenon and engineering, data are often used in the mathematical model development. There are also some mathematical models which were established from mathematical point of view, whose results are thought to be physically meaningful and to describe the nature. This can be considered the power of mathematics, which can not only help develop models using induction, i.e., from data, but also can make discoveries from pure mathematical logic. This book emphasizes the former.

This modeling book has a characteristics of interdisciplinary mathematics: elementary mathematics, calculus, linear algebra, statistics, big data, computer programming, meteorology, oceanography, and other application areas. Most existing mathematical modeling books are built on differential equation models, either ordinary differential equation or partial differential equation, and thus involve techniques of solving differential equations, either analytically or numerically. Those books are for senior or graduate levels of math majors, physics or engineering majors. This book is different and requires no prerequisite knowledge of differential equations and focuses on the needs of modern life's mathematical modeling using computer programming, such as linear regression, big data and Brownian motion. Thus, this book can be used by a mathematical modeling class for students in their second to fourth college year, majoring in math, physics, geoscience, engineering, computer science, and more. This book emphasizes the results finding and interpretation, although the model development method is also described.

Another feature of this book is to show students how to write short proposals and project reports based on mathematical modeling approaches.

–By SS in San Diego, January 2016

GLOSSARY

DD Calculus	Descartes' direct calculus
D[f,a]	Derivative of a function f at its indecent variable equal to a, which is the same as $f'(a)$
I[f,a,b]	Integral of a function f from a to b, which is the same as $\in_a^b f(x)dx$

Advanced Modern Mathematical Modeling
First Edition.

PART I

–DIMENSIONAL ANALYSIS
–LINEAR REGRESSION
–LINEAR ALGEBRA
–DATA MATRICES

CHAPTER 1

DIMENSIONAL ANALYSIS –A SHORTCUT TO OBTAIN A MATHEMATICAL MODEL FOR THE LAWS OF NATURE

This chapter shows a way to discover a mathematical model for a phenomenon using dimensional analysis, which requires the two sides of an equation to have the same dimension or units.

—Summary

1.1 Dimension and units

Length is called the dimension of a line, which is denoted by L. Length can be measured in SI Units: meter, or Imperial Units: feet. The SI is for French words "Systeme international d'unites", i.e., the International System of Units. SI system is also known as the metric system. Its commonly used length units are $m, dm, cm, mm, \mu m, km$; time units: $sec, \mu s$; and mass units: g, kg. The corresponding imperial system are $feet,\ lb,\ sec$. The imperial units is a British system. The SI system was published in 1960, and is now the most popular units system used in science and engineering around the world. The United States is the only major country that is still using the imperial units in engineering, but most science publications in the US have adopted the metric system. The United Kingdom had adopted the metric system in the 1960s.

Systematic use of units is very important. Misuse can can have serious consequences. On September 30, 1999, CNN reported that NASA lost a $125 million Mars orbiter because an engineering team mixed the two unit systems.
`http://www.cnn.com/TECH/space/9909/30/mars.metric.02/`.

Advanced Modern Mathematical Modeling
First Edition.
By Samuel Shen Copyright © 2016 John Wiley & Sons, Inc.

The units originated with culture, but nature laws should be independent of units. $F = ma$ works for both imperial and metric systems. Thus, it is critical that a law of nature is expressed in a single units system, not mixed systems.

1.2 Fundamental physics dimensions: $LMT\theta I$-class

The fundamental dimensions of physics are the five listed in Table 1.1.

Table 1.1 Fundamental dimensions: $LMT\Theta I$-class

Notation	Meaning	Dimension	Units
$[l]$	Length	L	m
$[m]$	Mass	M	kg
$[t]$	Time	T	sec or s
$[\theta]$	Temperature	Θ	$^\circ K$
$[I]$	Electric current (i.e., the flow flux of electric charge)	I	Amp or A

The dimensions of most other physical quantities can be derived from the above five. For example, speed is the displacement in a unit time and has is dimension LT^{-1}. Table 2 shows dimensions of the commonly used physical quantities.

EXAMPLE 1.1

Dimensional analysis of potential energy: The potential energy formula is

$$E = mgh. \tag{1.1}$$

Thus

$$[E] = [m][g][h] = M(LT^{-2})L = ML^2T^{-2}. \tag{1.2}$$

The last expression can be further organized into $M(LT^{-1})^2$, hence mass times speed squared, which has a clear physical meaning: kinetic energy $(1/2)mv^2$. This simple analysis links the potential energy and kinetic energy, and helps one to think about the conversion of potential energy into kinetic energy, such as the free fall of an iron ball, and vice versa.

Table 2 lists the dimensions of a few commonly used physical quantities.

EXAMPLE 1.2

Dimensional analysis of π: The constant π is a ratio of circumference to diameter $\pi = C/D$ for any circle. Thus

$$[\pi] = L/L = 1 \tag{1.3}$$

is dimensionless. π measures the angle of 180° is thus also dimensionless. Any angle can be measured by π or degree and is thus dimensionless, i.e., non-dimensional. The trigonometric functions, logarithmic functions, and exponential functions can only be applied to dimensionless quantities, such as 0.5π, or 2.3, or 1.0. These are pure numbers, but can also be regarded as radians, measuring an angle. However, radian is not

Table 1.2 Dimensions of derived physical quantities:

	Meaning	Dimension	SI Units
$[v]$	Velocity	LT^{-1}	m/s
$[a]$	Acceleration	LT^{-2}	m/s^2
$[F]$	Force (F=ma)	MLT^{-2}	$N = 1.0kg \cdot m/s^2$
$[\rho]$	Mass density	ML^{-3}	kg/m^3
$[p]$	Pressure (force per area)	$MLT^{-2}L^{-2} = ML^{-1}T^{-2}$	$Pa = N/m^2$
$[E]$	Energy	ML^2T^{-2}	$Joule = 1.0N \cdot m$
$[S]$	Entropy (energy per K)	$ML^2T^{-2}\Theta^{-1}$	$W/°K$
$[Q]$	Electric charge	IT	$C = 1.0A \cdot s$
$[E]$	Electric field (force per C)	$NC^{-1} = MLT^{-3}I^{-1}$	v/m
$[B]$	Magnetic field	$N(IL)^{-1} = MT^{-2}I^{-1}$	$T = 1.0kg/(As^2)$
$[\phi]$	Angle	1(dimensionless)	$radian$

a dimension. Of course, the range of trigonometric functions, logarithmic functions is also dimensionless. In the expressions $y = \sin x, y = \ln x, y = \exp(x)$, both x and y are dimensionless. In $\sin(\pi/6) = 0.5$, $\pi/6$ radian is considered dimensionless since radian is dimensionless, and 0.5 is also dimensionless.

Because of this common dimensionless feature of trigonometric functions, logarithmic functions and exponential functions, one may think that these functions should be related. Yes, they are. The exponential function and trigonometric functions are related by

$$e^{i\phi} = \cos\phi + i\sin\phi. \tag{1.4}$$

This is usually called Euler's formula (Leonhard Euler, 1707-1783, Swiss mathematician), illustrated by Fig. 1.1. Physics Nobel laureate Richard Feynman called Euler's equation“the most remarkable formula in mathematics.” This equation can help express numerous physical properties, such as wave function in quantum mechanics, homogeneity in universe, water waves, and alternative electric current.

The length of the arc of an interior angle θ and radius r is

$$s = \theta r. \tag{1.5}$$

The dimension of the above equation is

$$[s] = [\theta][r], \tag{1.6}$$

which is

$$L = [\theta]L. \tag{1.7}$$

Hence, $[\theta] = 1$ is dimensionless. This is another way to illustrate that angle is dimensionless, although we customarily use radian or degree to measure an angle. Neither radian nor degree for an angle should be considered dimensional.

The logarithmic function is the inverse function of exponential function. Other trigonometric functions can be derived from cosine and sine functions.

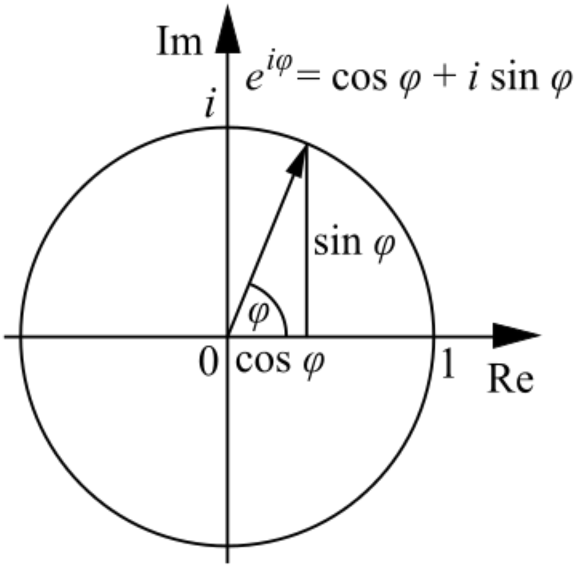


Figure 1.1 Euler's formula.

EXAMPLE 1.3

Dimensional analysis of pressure: The pressure's dimension is $ML^{-1}T^{-2}$ from Table 1.2 and can be re-written as $M(LT^{-1})^2L^{-3}$. Because $M(LT^{-1})^2$ is kinetic energy, $M(LT^{-1})^2L^{-3}$ is thus the kinetic energy per unit volume. This is the definition of pressure from the thermodynamic point of view. An ideal gas' pressure on its container wall is measured by the strength of the gas' kinetic energy per volume.

Thus, the rearrangement of different dimensions can result in very interesting and profound laws of nature. Dimensional analysis provides a powerful tool for discovery. We may say that dimensional analysis is a shortcut for discovery and can simplify experiments to lead to many useful mathematical formulas for physics and nature in general.

1.3 Relationships between magnetic and electric fields

From Table 1.2, the dimension of electric field is $MLT^{-3}I^{-1}$ which can be re-written as $(MT^{-2}I^{-1})(LT^{-1})$, whose first part is magnetic field and second part is velocity. When a conductor moves inside a magnetic field and cut the magnetic lines of force, an electric field can be felt because of the resistance exerted on the conductor, consequently, an electric current is generated and flows through the conductor (see Fig. 1.2). This is the principle of a power generator. Thus, dimensional analysis helps identify relevant physical quantities and can aid us to find new laws of physics, i.e., mathematical models for the physical quantities.

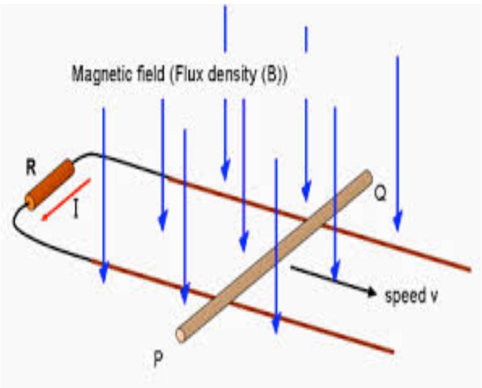


Figure 1.2 Generation of electric current when a conductor moves through a magnetic field and cuts the magnetic lines of force.

In general, a physical quantity X can be written as

$$X = \alpha l^a t^b m^c \theta^d I^e, \tag{1.8}$$

where α is a dimensionless constant. The dimensional analysis equation

$$[X] = \alpha [l]^a [t]^b [m]^c [\theta]^d [I]^e = \alpha L^a T^b M^c \Theta^d I^e \tag{1.9}$$

leads to linear equations for the exponents a, b, c, d, e. Solving these equations, one can obtain values of a, b, c, d, e. Equation (1.8) is then a mathematical model of a physics law, such as $(1/2)gt^2$ for the travelled distance of a free-fall body. Equation (1.9) is a special case of the general Buckingham's Π-theorem, which can be found in more detailed dimensional analysis books (Barenblatt 1987).

Unfortunately, dimensional analysis still cannot determine the value of α, which can be determined by an experiment or other mathematical approaches.

1.4 Dimensional analysis for a simple pendulum and calculation of the pendulum period

Simple pendulum clocks are based on the simple pendulum mechanism. For a clock, its most important function is to record time via the period of the pendulum, which is given by the following formula

$$T = 2\pi\sqrt{l/g} \tag{1.10}$$

where l is the length of the string and g is the gravitational constant. This formula can be derived using many methods, including an approach of the second order ordinary differential equation, which is beyond the scope of this book since most students in this class have not taken the ordinary differential equations course. Here we provide a simple approach via dimensional analysis. From the pendulum setup, it is reasonable to assume that the period is determined by mass of the pendulum, the length of the string, and Earth gravity. The Earth gravity might not be obvious, but one can think of an extreme environment: outpace of zero gravity, where the pendulum will not

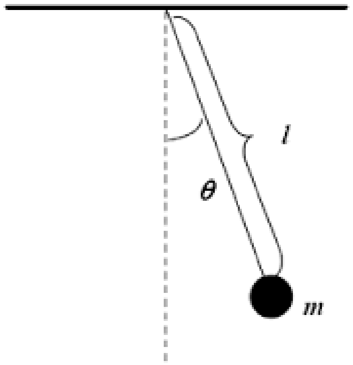


Figure 1.3 Simple pendulum of mass m and length l.

oscillate because of absence of gravity. The period is thus infinity. Similarly one may reasonably conclude that the same pendulum oscillates slower on moon than on Earth.

Thus, we can assume that the pendulum's period depends on mass, length and gravity, written in the following form:

$$T = \alpha m^a l^b g^c \tag{1.11}$$

with the exponent a, b, c to be determined by using the five fundamental dimensions:

$$[T] = [\alpha][m]^a[l]^b[g]^c = M^a L^b (LT^{-2})^c = M^a L^{b+c} T^{-2c}. \tag{1.12}$$

This implies

$$a = 0, \tag{1.13}$$

$$b + c = 0, \tag{1.14}$$

$$-2c = 1. \tag{1.15}$$

These equations have the following solutions

$$c = -1/2, b = 1/2, a = 0. \tag{1.16}$$

The period is proportional to $m^0 l^{1/2} g^{-1/2}$, or

$$T = \alpha m^0 l^{1/2} g^{-1/2} = \alpha\sqrt{l/g}. \tag{1.17}$$

An experiment was conducted in classroom with a string's length equal to 0.88 meters. Two periods were observed with time equal to 3.75 seconds. Substitute this into the above equation:

$$3.75 = 2 \times \alpha\sqrt{0.88/9.8} = 0.60\alpha, \tag{1.18}$$

$$\alpha = 3.75/0.60 = 6.25 = 1.99\pi =\approx 2\pi. \tag{1.19}$$

This is an easy experiment. Since the motion is relatively slow when the string is long enough, with smartphone stopwatch, it is fairly easy to record the time of two or three periods. One can improve the experimental results by making many experiments and use the average results as the final value for α.

The free-fall experiment is more difficult because the time is very short. Thus Galileo designed the experiment of a ball rolling down an inclined plate shown in Fig. 0.1.

1.5 Shock wave radius of a nuclear explosion

The instantaneous energy release from a nuclear explosion causes a shock wave, whose inside pressure is thousands times greater than outside. This pressure difference can push down trees and structures, and tear apart all kinds of objects. The shock wave may be assumed to be spherical and has radius R at t time after the explosion. Given the nuclear energy E, calculate the shock wave radius as a function of time, and hence predict the shock wave's arrival time and prepare for protection.

Shock wave occurring in atmosphere due to the supersonic compression of the air from one side so that the air mass from the side accumulates, cannot escape, builds pressure, develops a large pressure difference with the other side, and hence forms a shock. Two critical elements here are supersonic push and compressible air. Thus, shock wave radius should be related to density ρ of a compressible air, total energy E, and time t. Because air is light, gravity can be negligible. Thus, we assume the following

$$R = \alpha E^a \rho^b t^c. \tag{1.20}$$

The dimension of the above equation is

$$[R] = [\alpha][E]^a[\rho]^b[t]^c, \tag{1.21}$$

which leads to

$$L = 1 \times (ML^2T^{-2})^a(ML^{-3})^bT^c = M^{a+b}L^{2a-3b}T^{-2a+c}. \tag{1.22}$$

The exponents of both sides of this equation should be equal:

$$a + b = 0, \tag{1.23}$$

$$2a - 3b = 1, \tag{1.24}$$

$$-2a + c = 0. \tag{1.25}$$

These three equations have a unique solution

$$a = 1/5, b = -1/5, c = 2/5. \tag{1.26}$$

Therefore,

$$R = \alpha E^{1/5}\rho^{-1/5}t^{2/5}. \tag{1.27}$$

or

$$R = \alpha\left(\frac{Et^2}{\rho}\right)^{1/5}. \tag{1.28}$$

This makes Et^2 a very special term, which is the fifth power of density times length according to dimension equality, meaning the density of the air behind the shock.

Another expression of the shock radius is

$$R = \alpha\left(\frac{E}{\rho}\right)^{1/5}t^{2/5}. \tag{1.29}$$

The log-plot of this $T - t$ relationship is a straight line with slope 2/5:

$$\ln R = \ln\alpha + \ln\left(\frac{E}{\rho}\right)^{1/5} + \frac{2}{5}\ln t. \tag{1.30}$$

Any of the above three formulas can be used to predict the position of the shock wave for a given time, if α is known. Yet, it is not easy to evaluate this α by an experiment since such an experiment is too expensive. By solving another mathematical model, Cambridge University fluid dynamicist G. I Taylor (1886-1975) estimated that $\alpha = 1.0$.

Still another way of writing the above equation is

$$E = \frac{R^5 \rho}{t^2}. \tag{1.31}$$

This allows one to estimate the power of an nuclear bomb using news reports on the the shock arrival time at a given location.

If a nuclear bomb test is made underground, seismograph can measure R and t. With the known Earth crest's density, one can then estimate the bomb's power. There are 500 seismograph stations distributed around the world to detect ground-shaking incidents, including earthquakes and nuclear bombs.

One can use similar model to estimate the shock waves caused by supernova explosions (Exploring the X-ray Universe, Seward and Charles 2010).

EXERCISES

1.1 Make a dimensional analysis for the Newton's second law of motion: $F = ma$.

1.2 Design an experiment to demonstrate Newton's second law of motion: $F = ma$, when assuming the mass does not change and observing the data of acceleration and force.

1.3 Make a dimensional analysis for the gravitational force F_g based on law of universal gravitation shown in Fig. 1.4.

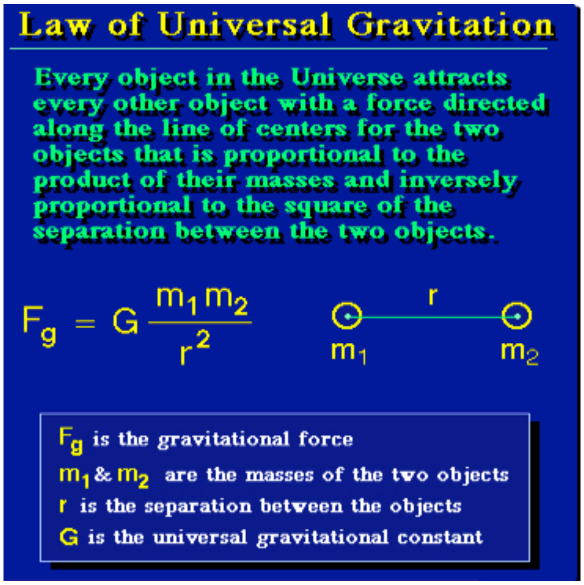
1.4 Dimensional analysis with experimental data.

a) Make a dimensional analysis for the velocity v and distance h of a free-fall body of mass m.

b) Perform the experiments of free-fall using a coin or any heavy metal or a stone to determine the dimensionless constant for distance. This is a difficult experiment since it happens really fast, and it is very hard to record time.

c) Change the free-fall experiment to free-roll experiment as Galileo did (see the preface of this book). Place a ball on an inclined plate and let it roll down by gravity. The gravity along the plate is now reduced to $g \sin \phi$ where ϕ is the angle between the plate and the flat flow (ideally the tangent plane perpendicular to the Earth's radius). The experiment is now easier since it is easier to record the time. However, the inclined angle should not be too small, which will make friction force non-negligible. Again, one can achieve better accuracy when repeating the experiment many times and using the average result.

References and Additional Reading Materials



Law of Universal Gravitation

Every object in the Universe attracts every other object with a force directed along the line of centers for the two objects that is proportional to the product of their masses and inversely proportional to the square of the separation between the two objects.

$$F_g = G\frac{m_1 m_2}{r^2}$$

F_g is the gravitational force

m_1 & m_2 are the masses of the two objects

r is the separation between the objects

G is the universal gravitational constant

Figure 1.4 Law of universal gravitation.

R1.1 G.I. Barenblatt, 1987: Dimensional Analysis, Gordon and Breach Science Publishers, New York, 354pp.

R1.2 F.D. Seward and P.A. Charles, 2010: Exploring the X-rays Universe, 2nd ed., Cambridge University Press, New York, 372pp.

CHAPTER 2

BASICS OF R PROGRAMMING

It is popular in today's mathematical modeling books to use computing tools for complex and tedious algebras so that students can focus on correct usage of the mathematical tools with accurate statement of assumptions and precise interpretation of the results. Among many software packages used in climate community, R's popularity has dramatically increased in the last a few years due to its enormous power of handling big data. We thus choose to include the basics of R for this book. A student who has mastered the R examples used in this book should have sufficient skills to develop R projects independently.

2.1 Download and install R software package

For Windows users, visit the website
`https://cran.r-project.org/bin/windows/base/`
to find the instructions of R program download and installations.

For Mac users, visit
`https://cran.r-project.org/bin/macosx/`

If you experience difficulties, please refer to online resources, Google or Youtube. A recent 3-minute Youtube instruction for R installation for Windows can be found from the following link:

https://www.youtube.com/watch?v=Ohnk9hcxf9M

The same author also has a youtube instruction about R installation for Mac (2 minutes):
https://www.youtube.com/watch?v=uxuuWXU-7UQ

One can choose to install R-Studio instead. Then visit
https://www.rstudio.com/products/rstudio/download/
This site allows to choose Windows, or Mac OS, or Unix.

One can use either R or R-Studio, or both, depending on his interest.

For details about the publicly open access to R-Project, visit
https://www.r-project.org/

The beginners of R would find it very difficult to navigate through this official, formal, detailed, and massive R-Project documentation to learn the program. Fortunately, many excellent tutorials for a quick learn of R programming are available online and in Youtube. One can google around and find a couple of preferred tutorials.

The following section provides R basics useful to this book.

2.2 R Tutorial

2.2.1 R as a smart calculator

R can be used like a smart calculator that allows fancier calculations than those done on regular calculators.

```
1+4
[1] 5
2+pi/4-0.8
[1] 1.985398
x<-1
y<-2
z<-4
t<-2*x^y-z
t
[1] -2
u=2        # "=" sign and "<-" is almost equivalent
v=3        # The text behind the "#" sign is comments
u+v
[1] 5
sin(u*v)    # u*v = 6 is considered radian
[1] -0.2794155
```

2.2.2 Write a function in R

The function command is
`name <- function(var1, var2, ...) expression of the function.`
For example,

```
square <- function(x) x*x
square(4)
[1] 16
fctn <- function(x,y,z) x+y-z/2
fctn(1,2,3)
[1] 1.5
```

2.2.3 Plot with R

R can can plot all kinds of curves, surfaces, statistical plots, and maps. Below are a few examples. For adding labels, ticks, color, and other features to a plot, one can google R plot and find the commands to properly include the desired features.

```
plot(sin, -pi, 2*pi)   #plot the curve of y=sin(x) from -pi to 2 pi
square <- function(x) x*x   #Define a function
plot(square, -3,2)   # Plot the defined function
fctn(1,2,3)
[1] 1.5
 ## Plot a 3D surface
x <- seq(-1, 1, length=100)
Z <- outer(x, x, function(x, y)sqrt(1-x^2-y^2))
#outer (x,y, function) is outer product
persp(x=x, y=x, z=Z, theta=310)
# yields a 3D surface with perspective angle 310 deg
```

2.2.4 Symbolic calculations by R

People used to think that R can only handle numbers. Actually R can do symbolic calculations, such as finding a derivative. However, up to now R is not the best symbolic calculation tool. One can use WolframAlpha, SymPy, and Yacas for free or use the paid software package Maple or Mathematica. Google symbolic calculation for calculus to find a long list of symbolic calculation software packages, such as `https://en.wikipedia.org/wiki/List_of_computer_algebra_systems`.

```
D(expression(x^2,'x'), 'x')
# Take derivative of x^2 and the answer is 2x
2 * x
fx= expression(x^2,'x')  #assign a function
D(fx,'x') #differentiate the function with result below
2 * x
fx= expression(x^2*sin(x),'x')
#Change the expression and use the same derivative command
D(fx,'x')
2 * x * sin(x) + x^2 * cos(x)
```

```
 fxy = expression(x^2+y^2, 'x','y')
#One can define a function of 2 or more variables
 fxy #This gives the expression of the function in terms of x and y
expression(x^2 + y^2, "x", "y")
D(fxy,'x') #This gives the partial derivative with respect to x: 2 * x
D(fxy,'y') #This gives the partial derivative with respect to y: 2 * y
square = function(x) x^2
integrate (square, 0,1)
#Integrate x^2 from 0 to 1 equals to 1/3 with details below
0.3333333 with absolute error < 3.7e-15
integrate(cos,0,pi/2)
#Integrate cos(x) from 0 to pi/2 equals to 1 with details below
1 with absolute error < 1.1e-14
```

2.2.5 Vectors and matrices

R can handle all kinds of operations vectors and matrices.

```
c(1,6,3,pi,-3) #c() gives a vector and is considered a 4X1 column vector
[1]  1.000000  6.000000  3.000000  3.141593 -3.000000
seq(2,6) #Generate a sequence from 2 to 6
[1] 2 3 4 5 6
seq(1,10,2) # Generate a sequence from 1 to 10 with 2 increment
[1] 1 3 5 7 9
x=c(1,-1,1,-1)
x+1 #1 is added to each element of x
[1] 2 0 2 0
2*x #2 multiplies each element of x
[1]  2 -2  2 -2
x/2 # Each element of x is divided by 2
[1]  0.5 -0.5  0.5 -0.5
y=seq(1,4)
x*y  # This multiplication * multiples each pair of elements
[1]  1 -2  3 -4
x%*%y #This is the dot product of two vectors and yields
     [,1]
[1,]   -2
t(x)  # Transforms x into a row 1X4 vector
     [,1] [,2] [,3] [,4]
[1,]    1   -1    1   -1
t(x)%*%y #This is equivalent to dot product and forms 1X1 matrix
     [,1]
[1,]   -2
> x%*%t(y) #This column times row yields a 4X4 matrix
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]   -1   -2   -3   -4
[3,]    1    2    3    4
[4,]   -1   -2   -3   -4
```

```
my=matrix(y,2,2)
#Convert a vector into a matrix of the same number of elements
#The matrix elements go by column, first column, second, etc
     [,1] [,2]
[1,]    1    3
[2,]    2    4
dim(my)  #find dimensions of a matrix
[1] 2 2
as.vector(my) #Convert a matrix to a vector, again via columns
[1] 1 2 3 4
mx*my #multiplication between each pair of elements
     [,1] [,2]
[1,]    1    3
[2,]   -2   -4
mx/my #division between each pair of elements
     [,1]       [,2]
[1,]  1.0  0.3333333
[2,] -0.5 -0.2500000
mx-2*my
     [,1] [,2]
[1,]   -1   -5
[2,]   -5   -9
mx%*%my #This is the real matrix multiplication in matrix theory
     [,1] [,2]
[1,]    3    7
[2,]   -3   -7
det(my) #determinant
[1] -2
myinv = solve(my) #yields the inverse of a matrix
> myinv
     [,1] [,2]
[1,]   -2  1.5
[2,]    1 -0.5
> myinv%*%my #verifies the inverse of a matrix
     [,1] [,2]
[1,]    1    0
[2,]    0    1
diag(my) #yields the diagonal vector of a matrix
[1] 1 4
myeig=eigen(my) #yields eigenvalues and unit eigenvectors
myeig
$values
[1]  5.3722813 -0.3722813
$vectors
           [,1]       [,2]
[1,] -0.5657675 -0.9093767
[2,] -0.8245648  0.4159736
mysvd = svd(my) #SVD decomposition of a matrix M=UDV’
            #SVD can be done for a rectangular matrix of mXn
```

```
mysvd
$d
[1] 5.4649857 0.3659662
$u
           [,1]       [,2]
[1,] -0.5760484 -0.8174156
[2,] -0.8174156  0.5760484
$v
           [,1]       [,2]
[1,] -0.4045536  0.9145143
[2,] -0.9145143 -0.4045536

ysol=solve(my,c(1,3))
#solve linear equations matrix %*% x = b
ysol  #solve(matrix, b)
[1]  2.5 -0.5
my%*%ysol #verifies the solution
     [,1]
[1,]    1
[2,]    3
```

2.2.6 Statistics

R was originally designed by statisticians for doing statistics. Thus, R has a comprehensive set of statistics functions. This sub-section gives a few basic commands. More will be described in the statistical modeling chapters.

```
x=rnorm(10) #generate 10 normally distributed numbers
x
 [1]  2.8322260 -1.2187118  0.4690320 -0.2112469  0.1870511
 [6]  0.2275427 -1.2619005  0.2855896  1.7492474 -0.1640900
 mean(x)
[1] 0.289474
var(x)
[1] 1.531215
sd(x)
[1] 1.237423
median(x)
[1] 0.2072969
quantile(x)
        0%        25%        50%        75%       100%
-1.2619005 -0.1994577  0.2072969  0.4231714  2.8322260
range(x) #yields the min and max of x
[1] -1.261900  2.832226
 max(x)
[1] 2.832226

boxplot(x) #yields the box plot of x
w=rnorm(1000)
```

```
summary(rnorm(12)) #statistical summary of the data sequence
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.9250 -0.6068  0.3366  0.2309  1.1840  2.5750

hist(w)
#yields the histogram of 1000 random numbers by normal distribution
```

2.3 Online Tutorials

2.3.1 Youtube tutorial: for true beginners

This is a very good and slow paced 22 minutes youtube tutorial: Chapter 1. An Introduction to R
https://www.youtube.com/watch?v=suVFuGET-0U

2.3.2 Youtube tutorial: for some basic statistical summaries

This is a 9 minutes tutorial by Layth Alwan.
https://www.youtube.com/watch?v=XjOZQN-Nre4

2.3.3 Youtube tutorial: Input data by reading a csv file into R

An excel file can be saved as csv file: xxxx.csv. This 15 minutes youtube shows how to read a csv file into R by Layth Alwan. He also shows linear regression.
https://www.youtube.com/watch?v=QkE8cp0B9gg

R can input all kinds of data files, including xlsx, netCDF, fortran data, and sas data. Some commands are below. One can google to find proper data reading command for your particular data format.

```
mydata <- read.csv("mydata.csv")
# read csv file named "my data.csv"

mydata <- read.table("mydata.txt")
# read text file named "my data.txt"

library(gdata)                    # load gdata package
mydata = read.xls("mydata.xls")   # read an excel file

library(foreign)                  # load the foreign package
mydata = read.mtp("mydata.mtp")   # read from .mtp file

library(foreign)                  # load the foreign package
mydata = read.spss("myfile", to.data.frame=TRUE)

ff <- tempfile()
cat(file = ff, "123456", "987654", sep = "\n")
read.fortran(ff, c("F2.1","F2.0","I2")) #read a fortan file
```

```
library(ncdf)
ncin <- open.ncdf(ncfname)  # open a NetCDF file
lon <- get.var.ncdf(ncin, "lon") #read a netCDF file into R
```

Some libraries are not in the R project anymore. For example,

```
library(ncdf) #The following error message pops up
Error in library(ncdf) : there is no package called ncdf
```

One can then google r data reading netcdf R-project and go to the R-project website. The following can be found.

```
Package ncdf was removed from the CRAN repository.
Formerly available versions can be obtained from the archive.
Archived on 2016-01-11: use ’RNetCDF’ or ’ncdf4’ instead.
```

This means that one should use RNetCDF, which can be downloaded from internet. Thus, if a library gives an error message, then google the library package, download and install the package, and finally read the data of the particular format.

References and Additional Reading Materials

R2.1 R tutorial by Steve Jost, De Paul University,

`http://facweb.cs.depaul.edu/sjost/csc423/`

R2.2 R tutorials by William B. King, Coastal Carolina University,

`http://ww2.coastal.edu/kingw/statistics/R-tutorials/`

CHAPTER 3

LINEAR MODELS BY REGRESSION

3.1 Introduction to a linear model

In our routine life, we often forecast something based on data using a linear model, such as "If its economy grows at this speed, China's GDP will surpass US in 30 years." This kind of forecast is usually based on a linear model whose slope and an initial point are calculated by observed data. In this US-China GDP case, the forecast is based on the double digit growth of China's GDP in the period of 2000-2012. The critical parameter of a linear model is the growth rate, or called the rate of change, or slope, or derivative. The linear model is thus a straight line model with an initial point, also called y-intercept, and a slope. The mathematical expression is

$$y = a + bx \tag{3.1}$$

where a is the intercept and b is the slope. The linear modeling process is to estimate a and b when data are given and to make inferences, conclusions and discussion based on the estimates.

Figure 3.1 shows a linear model between the heating degree day (HDD) and the energy consumption needed. HDD of a day in the US is defined as 65°F minus the average temperature of of the day, which is usually defined as the mean of the daily maximum temperature T_{max} and the daily minimum temperature T_{min}

$$HDD = 65 - \frac{T_{max} + T_{min}}{2} \geq 0 \tag{3.2}$$

and $HDD = 0$ is the above formula yields negative value. The monthly cumulative HDD is often used to forecast the heating energy needed for a facility, such as a university campus.

The linear model in Figure 3.1 is based on 14 data points. The linear model is

$$y = 3.317x + 53.505, \tag{3.3}$$

where x is HDD and y is KWh. The linear model can be used for predicting energy usesage for any given HDD value. If the prediction is outside the range of the x data, then it is called extrapolation, such as predicting the energy consumption for an extremely cold month HDD=500. The linear model prediction is

$$3.317 \times 500 + 53.505 = 1,712 \; [KWh]. \tag{3.4}$$

If the prediction is within the range of the x data, but not on the data, then the prediction is called interpolation, such as predicting the energy consumption for HDD= 210, whose linear model prediction is

$$3.317 \times 210 + 53.505 = 750 \; [KWh]. \tag{3.5}$$

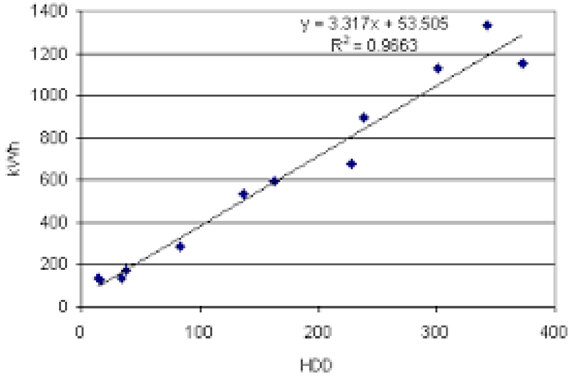


Figure 3.1 Heating energy (KWh) predicted by a linear model with respect to the heating degree day HDD.

In general, a linear model is rigorously written as

$$y = a + bx + \epsilon \tag{3.6}$$

where a is called intercept, b is called trend, or, slope, and ϵ is called error, y is the response variable, or dependent variable, and x is the explanatory variable, or independent variable.

The estimated linear model is

$$\hat{y} = \hat{a} + \hat{b}x, \tag{3.7}$$

where $\hat{a}$ and $\hat{b}$ can be computed by R when the data of dependent and independent variables are given. Namely, when one can plot a the points based on the (x, y) data, R can calculate a linear line that best fits the data with minimum square error. This is why regression, meaning returning to an object, a situation or a state, is also called the least square regression. English word “regression” originated in 15th century from Latin “regressionem”, meaning a going back, a return, according to etymology dictionary.

If there are only two data points, then the best fit is a perfect fit to the two points since any two points determine a line. The regression line will pass the two points , and the mean square error is zero.

If there are more two points which are not distributed on a line, then the least square residuals guarantees the best mid-line which leaves points on both sides of the line. The linear regression is a branch of statistics and studies the errors, trend, reliability of estimated values, statistical inferences, and extensions to multi-variables and nonlinear models. This book only includes the materials that using R to compute linear models and does not provide details of the error estimation and inferences.

3.2 Evaluate a linear model by regression using R

A linear model can be determined by R given data. Let us use an example to describe the linear modeling procedures by R: examine the linear trend of the global average annual mean land surface air temperature change from 1880 to 2014. This is based on data produced by James Hansen’s NASA climate research group:
`http://cdiac.ornl.gov/trends/temp/hansen/hansen.html`
I downloaded the land meteorological station data from
`http://cdiac.ornl.gov/ftp/trends/temp/hansen/gl_land.txt`
by clicking Firefox’s File drop down menu’s
`Save Page As ...`
I saved it into a directory of my own laptop computer

```
~/Desktop/MyDocs/teach/336MathModel-2016SP/
BookMathModeling2016/Notes4Book/ModCh3/gl_land.txt
```

The file name is
`gl_land.txt`,
the same as the original file name on the website. This file has a head and foot. To avoid complication, I manually deleted both head and foot and left the file with only the digital data. I saved this edited filed as
`gl_land_nohead.txt`

Read the txt data into R by

```
dtmean<-read.table("~/Desktop/MyDocs/teach/336MathModel-2016SP/
BookMathModeling2016/Notes4Book/ModCh3/gl_land_nohead.txt", header=F)
```

Because the txt file now has no head, in the reading command “head” is given value “false” F. The txt data are now read into the R environment as R data filed named
`dtmean`
Enter `dtmean` in R console and the data will show

```
> dtmean
      V1    V2     V3
1   1880 -0.43 -99.99
2   1881 -0.34 -99.99
```

```
3   1882 -0.28  -0.38
4   1883 -0.28  -0.39
5   1884 -0.57  -0.43
............
```

Use `dim(dtmean)` to check the dimension of the dataset.
With the above data reading preparation, the following

```
> dim(dtmean)
[1] 135   3
```

This is correct: 3 columns (the first column is year, the second the temperature, and the third the 5-year moving mean), and 135 rows meaning 135 years from 1880 to 2014.

Generate two vectors: one for the time ticks: year 1880 to 2014, and one for the temperature of each year.

```
yrtime<-dtmean[,1]
tmean<-dtmean[,2]
```

Use statistical summary to get some crude information on these two data vectors.

```
>summary(tmean)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.660000 -0.255000 -0.060000  0.007556  0.190000  0.910000
> summary(yrtime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1880    1914    1947    1947    1980    2014
```

These summaries seem right.

One can also make a boxplot and a histogram to see if the data are reasonable or if the data can show some important information.

The command `boxplot(tmean, main="Land Temp Anomalies")` generates a box plot shown in Figure 3.2.

The command

```
hist(tmean, main="Histogram: Land Temperature Anomalies",xlab="Temp anomalies")
```

generates a histogram in Figure 3.3.

We can easily visually view the data since these datasets are small.

```
>tmean
  [1] -0.43 -0.34 -0.28 -0.28 -0.57 -0.46 -0.57 -0.66 -0.40 -0.20 -0.57 -0.61 -0.52
 [14] -0.54 -0.43 -0.35 -0.32 -0.24 -0.42 -0.35 -0.17 -0.17 -0.36 -0.40 -0.52 -0.35
   ......
> yrtime
   [1] 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894 1895
  [17] 1896 1897 1898 1899 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911
   ......
```

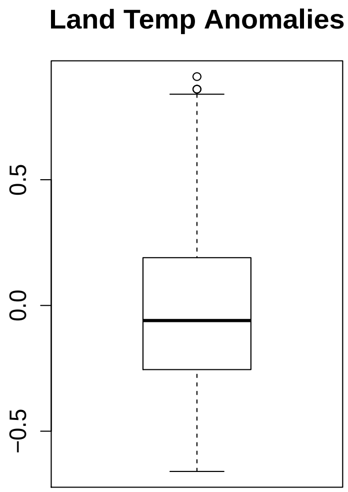


Figure 3.2 Boxplot of the global land average annual mean surface air temperature anomalies from land stations only: 1880-2014.

Implement the linear model using R

```
reg8014<-lm(tmean ~ yrtime)
```

Here "lm" means linear model. The first dataset "tmean" is the vertical axis and the second dataset "yrtime" is for the horizontal axis. The linear model's calculation results are placed in the file named "reg8014". Please pay special attention to the confusion positions of x-y data in this R command `lm(tmean ~ yrtime)` where the y-axis data is ahead of the x-axis data. This is opposite to the `plot(yrtime, tmean)` where the y-axis data is behind the x-axis data.

To see regression results, use command "summary(reg8014)"

```
> summary(reg8014)

Call:
lm(formula = tmean ~ yrtime)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45381 -0.12141  0.00266  0.10923  0.35180

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.720e+01  7.058e-01  -24.36   <2e-16 ***
yrtime       8.836e-03  3.625e-04   24.38   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

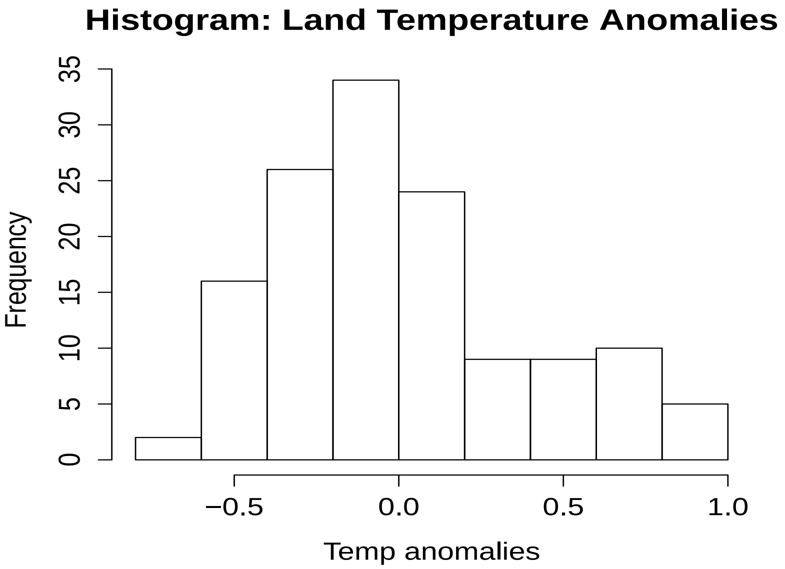


Figure 3.3 Histogram of the global land average annual mean surface air temperature anomalies from land stations only: 1880-2014.

```
Residual standard error: 0.1641 on 133 degrees of freedom
Multiple R-squared:  0.8171,      Adjusted R-squared:  0.8158
F-statistic: 594.3 on 1 and 133 DF,  p-value: < 2.2e-16
```

The most important information from the above results is the estimated linear model with intercept and trend:

$$tmean = -1.720e + 01 + 8.836e - 03 \times year \tag{3.8}$$

i.e.,

$$y = -17.2 + 0.008836x. \tag{3.9}$$

See Figure 3.4 for the data point positions and the linear regression line.

The command

```
plot(yrtime,tmean,xlab="Year",ylab="Land temperature",
main="Global Annual Mean Land Surface Air Temperature", type="o")
```

plots the 135 data points which are linked by a line. `type="o"` means linking all the data points by a line. Without `type="o"`, the plot will show only the 135 points.

The command

```
abline(reg8014, col="red")
```

adds the linear regression line onto the previous plot.

One can add text to the plot too. For example,

```
text(1930, 0.6, "Linear temp trend 0.88 oC per century", col="red",cex=1.2)
```

adds the text “Linear temp trend 0.88 oC per century” centered at the point (1930, 0.6).

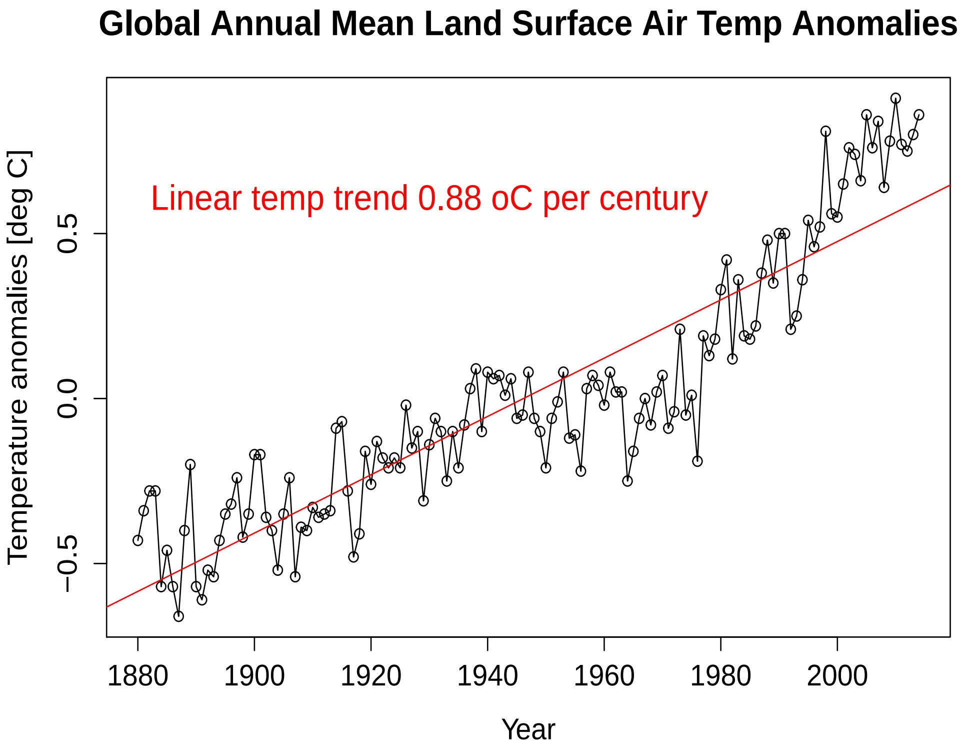


Figure 3.4 Global land average annual mean surface air temperature anomalies from land stations only: 1880-2014. The meteorological station temperature anomalies are computed with respect to the 1951-1980 mean, which is also called climatology.

In the linear model summary, the Std. Error indicates how reliable is the linear model. In our case, the error for intercept is 0.7, which is about 4% of the intercept value 17. It is very accurate. Thus, the one-side t-statistic is large and equal to -24.36 and the p-value is very small 2×10^{-16}. This implies that the intercept is significant even at 1% significance level and the intercept value is highly reliable in this linear model.

The standard error for the trend, i.e., the slope, is 0.0003625, which is only about 4% of the trend 0.008836 °C/year. It is also very accurate. The one-side t-statistic is 24.38, again very large. The p-value is 2×10^{-16} again very small. This implies that the trend is significant even at 1% significance level and the trend value is highly reliable in this linear model.

The residuals are the vertical differences between the data points and the linear regression line. There are 135 residuals, which form a data vector whose minimum is -0.45381 and maximum is 0.35180. We can use

```
residuals.lm(reg8014)
```

to show the values of all the residuals. This of course can also show which year the maximum and minimum residuals occurred. In our case, min occurred at 97th year

from 1880, i.e., 1976, and max at the 119th year from 1880, i.e., 1998. The first six months of 1998 were at the fading phase of a very strong El Nino, which often enhance positive anomalies of the land temperature. The first four months of 1976 were at the end of a strong La Nina, which often enhance negative anomalies of the land temperature. See the US NOAA Climate Prediction Center's website for the El Nino and La Nina monitoring:
`http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml`

The residual standard error in the linear model summary is defined as

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{135} \hat{\epsilon}_i^2}{135-2}}, \tag{3.10}$$

where $135-2$ is the degrees of freedom (dog) because two regression parameters a and b are estimated and hence two constraints are implemented onto the 135 data points, and $\hat{\epsilon}_i = D_i - \hat{y}_i$ is the residual with data D_i and the linear model value $\hat{y}_i = a + bx_i$.

From the above formula, we can see that the residual standard error is almost the same as the standard deviation of the residuals, whose unbiased estimator is calculated by the following formula:

$$SD_\epsilon = \sqrt{\frac{\sum_{i=1}^{135} \hat{\epsilon}_i^2}{135-1}}, \tag{3.11}$$

since the mean of the $\hat{\epsilon}_i (i = 1, ..., 135)$ is zero. This result of zero is expected since residuals are the differences of data and the regression line. By the English meaning of regression, the line should be in the middle, and hence the positive and negative residuals should be cancelled. However, this intuitive conclusion needs a mathematical proof, which is a topic in mathematical statistics and is not derived here.

Thus,

$$\hat{\sigma} = SD_\epsilon \sqrt{\frac{n-1}{n-K}}. \tag{3.12}$$

One can verify this formula by computing

```
rel <- residual.lm(reg8014)
sd(rel)*sqrt((135-1)/(135-2))
```

The output is `[1] 0.1641161`, which is the residual standard error $\hat{\sigma}$.

The multiple R-squared value in the summary is equal to $R^2 = 0.817$, which is quite large and indicates that all the data points are around a straight line, and hence supports the validity of a linear model. This R-squared value is the square of the correlation between `tmean` and `yrtime`. One can verify this R-squared value by `(cor(tmean, yrtime))^2`.

The adjusted R-squared value R_a^2 is related to R^2 by

$$R_a^2 = \frac{n-1}{n-K} R^2 - \frac{K-1}{n-K}, \tag{3.13}$$

where n is the total number of data points (n=135), K is the total number of constraints (K=2 because of the estimation of a and b: intercept and slope). The multiple

R-squared value 0.8171 in the linear model summary has little difference from the adjusted R-squared 0.8158 when the sample size is large (usually means more than 50), like our case of n=135. These two values measure how the points distribute around a straight line. When these values are close to zero, the linear model is invalid since the points are scattered around randomly or in a fast oscillation. Either the points are truly random or the data follow a nonlinear model, not a straight line linear model.

The F-statistic 594.3 in the linear model summary is used to test whether the slope is significantly different from zero using F-test in the analysis of variance (ANOVA). When F value is greater than 5, the slope may be regarded as significantly different from zero. However, in practice F-test for the non-zero slope is very sensitive, hence not very reliable. One should make his own conclusion on the non-zero slope, i.e., trend, from the regression line, the regression plot like Fig. 3.4 and the entire linear model summary.

3.3 Research level exploration

The linear trend from 1880-2014 is 0.88 °C per century from Fig. 3.4 based on Hansen's data. One may ask question: how does the trend change if looking at different time periods, e.g., 1880-1910 (a relatively flat period), 1880-1950 (1950 being before a short period of global cooling), 1880-1975 (1975 being in the middle of a global cooling period from 1960-1980), 1880-2000 (2000 being the beginning of a high plateau), and 1880-2014 (2014 being the most current)?

We can read each section of the data out, calculate a linear model, plot the linear model line, and print the trend value. We use one section (1880-1910) to describe the R computing code, as example to illustrate the entire procedure.

```
x8010=seq(1880,1910) #generate x-axis data
y8010=seq(1,31) #generate y-axis data positions
for (i in 1:31) {y8010[i]=dtmean[[i,2]]} #form y-data
reg8010<-lm(y8010 ~ x8010)
plot(yrtime,tmean,xlab="Year",ylab="Temperature anomalies [oC]",
main="Global Annual Mean Land Surface Air Temp", type="o")
#plot the main curve and data
reg8014<-lm(tmean ~ yrtime) #linear model again for all data
abline(reg8014, col="red")
text(1920, 0.90, "1880-2014 trend= 0.88 oC/100a", col="red",cex=1.0)
abline(reg8010, col="blue")
text(1920, 0.3, "1880-1910 trend= 0.32 oC/100a", col="blue",cex=1.0)
```

The above lines of R code can produce linear regression lines of 1880-2014 and 1880-1910. In the similar way, one can calculate and plot the linear models of the other three time periods. The final result with all the five linear models in five different periods of time is shown in Fig. 3.5.

The trends computed from the five different periods are in the range (0.32, 0.88)°C per century. If we wish to predict the temperature in 2030 using a linear model extrapolation, which linear model should we use? The 1880-2014 linear model yields 0.74°C while the 1880-1975 linear model leads to 0.39°C, about half of the former.

The temperature in 1880-2014 has a general increase trend, but can decrease from year to year or decrease in an extended period of time, such as 1960-1980. Linear

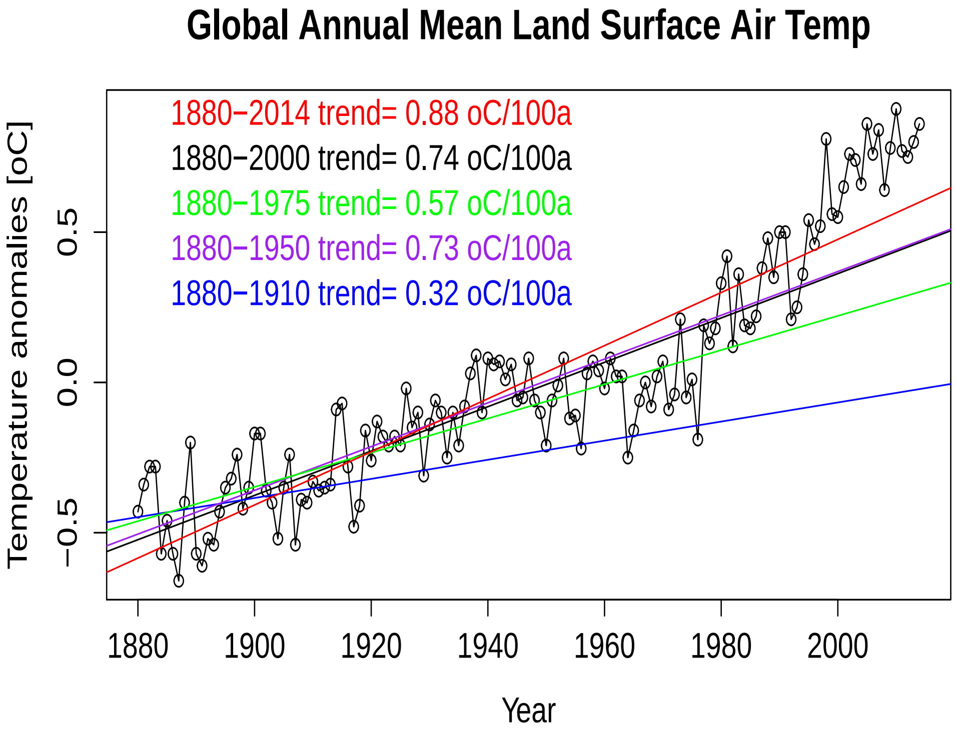


Figure 3.5 Trends of the global annual mean land surface air temperature anomalies in five different periods: 1880-2014, 1880-1910, 1880-1950, 1880-1975, 1880-2000, and 1880-2014.

models can give some first order prediction of the future temperature, but much uncertainty exists. One even cannot exclude the possibility of another cooling period like that of 1960-1980. More complex models and data are thus needed to make climate predictions for the future climate in the next decades or century. Therefore, this question of global climate extrapolation is not simple and is an extensive research topic in geoscience, mathematics, statistics and computing science.

EXERCISES

3.1 One of the most commonly used datasets of global average annual mean surface air temperature (SAT) anomalies (relative to 1951-1980 climatology period) from 1880-2014 was produced by James Hansens NASA research group
`http://cdiac.ornl.gov/trends/temp/hansen/hansen.html`
The anomalies are defined as the departure from 1951-1980s average, which is called climatology.

a) Read the annual temperature anomalies data for the entire globe including both ocean and land into R and compute the statistical summary of this dataset.

b) Make a box plot of the data.

c) Plot the histogram of the data.

d) Find the linear regression models T=a + b t of the data for the following periods: (i) 1880-2014, (ii) 1880-1910, (iii) 1880-1950, (iv) 1880-1975, and (v) 1880-2000. Find a, b values and put these values in a table. Plot the linear regression lines and the data time series on a single figure. Use different colors for the regression lines in the different period.

3.2 Another commonly used dataset of global average annual mean surface air temperature (SAT) anomalies (relative to 1961-1990 climatology period) from 1850-2014 was produced by Phillip Jones UK research group at University of East Anglia
`http://cdiac.ornl.gov/trends/temp/jonescru/jones.html`
The anomalies are defined as the departure from 1961-1990s average.This 30-year average is also called climatology [1]. Please do the following:

a) Read the global annual temperature anomalies data into R and compute the statistical summary of this dataset.

b) Make a box plot of the data.

c) Plot the histogram of the data.

d) Find the linear regression models T=a + b t of the data for the following periods: (i) 1850-2014, (ii) 1850-1910, (iii) 1850-1950, (iv) 1850-1975, and (v) 1850-2000. Find a, b values and put these values in a table. Plot the linear regression lines and the data time series on a single figure. Use different colors for the regression lines in the different period.

3.3 Make the similar linear model analysis for the January's average Tmin temperature from station Cuyamaca (32.9897°N, 116.5872°W) from 1951-2010. This station is at the eastern suburb of San Diego, USA. The station ID in the United States Historical Climatology Network (USHCN) is 042239. One can download the data from the website
`http://cdiac.ornl.gov/ftp/ushcn_v2.5_monthly/`
Select
`ushcn2014_FLs_52i_tmin.txt.gz`
The R-squared value is 0.01314, very small, indicating wide scattering of the data and the linear model is inappropriate for Cuyamaca's January Tmin temperature from 1951-2010. Use R to go through the linear model procedures and make a more comprehensive conclusion.

[1] Climatology is the mean and normal state of climate, but its calculation does not have a fixed way. Using 1961-1990 for the global climatology is the common practice now because the this period has most temperature and precipitation station data available to the public. The mean state of a climate is not uniquely defined in mathematics since the a new mean state may be reached due to long-term climate changes, such as the warming trend since 1850 to 2010. The climatology may be defined mathematically in deferent ways, including 30-year average, an intrinsic mode function with an average period equal to one year in Hilbert-Huang Transform, wavelet definition, or Fourier series definition (Shen et al. 2005).

References and Additional Reading Materials

R3.1 M. Maathuis (2015): R-regression tutorial and statistical theory by Marloes Maathuis, ETH Zurich, Switzerland:
`http://stat.ethz.ch/˜mmarloes/teaching/fall08/5-LinearRegression.pdf`

R3.2 K. Van Steen (2015): R-regression tutorial and statistical theory by Kritel Van Steen, Montefiore Institute, Belgium:

`http://www.montefiore.ulg.ac.be/˜kvansteen/GBIO0009-1`
`/ac20092010/Class8/Using%20R%20for%20linear%20regression.pdf`

R3.3 Climate Prediction Center of the United States (2015): Historical El Nino and La Nina: cold and warm episodes by season since 1950

`http://www.cpc.ncep.noaa.gov/products`
`/analysis_monitoring/ensostuff/ensoyears.shtml`

R3.4 Shen, S.S.P., T. Shu, N.E. Huang, Z. Wu, G.R. North, T.R. Karl, and D.R. Easterling (2005) HHT analysis of the nonlinear and non-stationary annual cycle of daily surface air temperature data. In Hilbert-Huang Transform and Its Applications, edited by N.E. Huang and S.S.P. Shen, World Scientific, Singapore, pp.187-210.