Advanced statistical analysis using R

— Factor analysis, SEM, applied regression models —

(Text for lecture of statistical analysis in health and medicine)

Graduate School of Health Sciences, Kobe Univ.
Prof. Minato Nakazawa
<minato-nakazawa@people.kobe-u.ac.jp>

2 June 2025

Revisions
0.1 Translated from Japanese (not completed yet), partly using Google gemini (May 30, 2025)
0.2 Revised (still not completed yet). (June 2, 2025)

Contents

Preparation to take this class

The aim of this class is to explain some of the popular but relatively advanced-level statistical data analysis methods. The goal is that the students can conduct such analysis by themselves, so that practical examples using statistical software are explained.

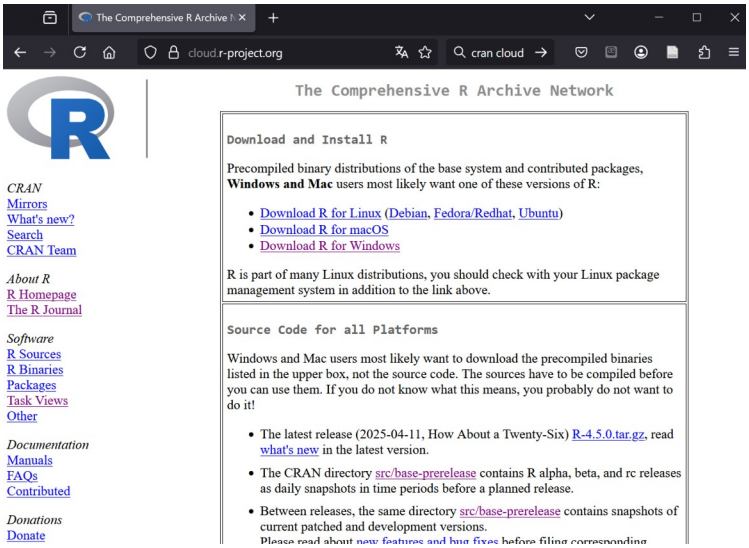
This class adresses advanced statistical analysis, so that I presume that basic knowledge on statistical analysis have been already learned. If you have any ambiguity or questions about such basic knowledge on statistics, please query to dictionary or textbook of statistics. For instance, Upton and Cook (2014) provides a good glossary.

There are many famous statistical analysis softwares such as SAS, SPSS and others, most of which are very expensive, but nowadays R is the best software to be used in this class, because (1) it's free software and thus anyway, anytime, anybody can use it in one's own computer, (2) almost all statistical methods are included or available as additional package, which can be installed from CRAN or bioconductor or GitHub web sites, (3) the results are so reliable as accepted by top journals such as *Nature*, *Science*, *Cell* and *Proceeding of National Sciences, USA*. R is implemented in any of Windows, MacOS, and Linux. Please install R into your own computer to use in the class.

To manipulate R via Graphical User's Interface (GUI), I recommend you to install RStudio, also free software. In another lecture on basic statistical analyses (Hokengaku-kyotsu-tokko IV, VIII), I have explained statistical analyses using EZR package, which is also GUI for R. However, EZR is menu-driven package and thus not suitable for this advanced statistics class.

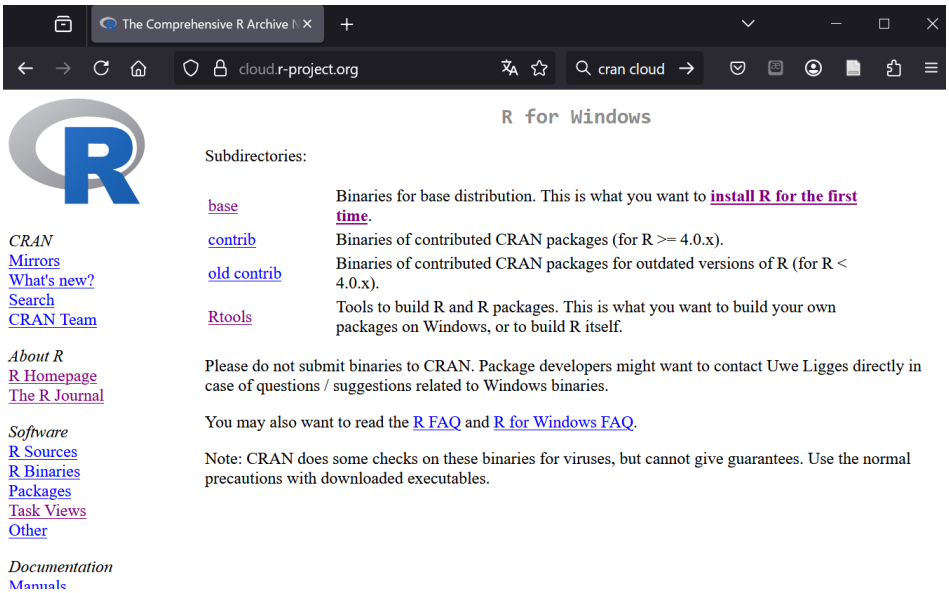
Recently, another free software, jamovi (https://www.jamovi.org/) is available to conduct some of the advanced statistical analyses. The jamovi is very easy to use, and reliable because the calculations are done by R background. Free textbook for jamovi, "Learning Statistics with jamovi" is also available in the web site (https://www.learnstatswithjamovi.com/). Therefore, if you plan to conduct some typical analyses, jamovi is also a good candidate.

Installation of R



You can download the most software related to R from CRAN (https://cloud.r-project.org). The screenshot of the top page of CRAN is shown above.

Depending on your computer's OS (Windows, MacOS or Linus), you must follow the suitable links. In the case of Windows, you see the screen below.



On 13 May 2025, the latest R version is R-4.5.0. If your OS is Microsoft Windows, I recommend you to download **R-4.5.0-win.exe** from the **[base]** link and R Tools 4.5 from **[Rtools]** link following the links and finally downloading **rtools45-6536-6492.exe**.

After downloading, you have to run R-4.5.0.exe as administrator. To do so, usually right-click the icon of R-4.5.0.exe and choose [Run as administrator] from context menu. You may choose manual setting as install option, then you should choose SDI. Other options are not so important and up to you.

I cannot fully explain how to install R in MacOS or Linux here. Please find suitable explanation in any web sites. It's not so difficult.

Installation of RStudio

The RStudio is available at https://posit.co/downloads/. Many famous developer of R packages join this project such as Dr. Yuhui Xie (known by animation and knitr) and Dr. Hadley Wickham (no need to mention, well known by tidyverse including ggplot2, plyr, dplyr and devtools.

For most users, click the button [DOWNLOAD] under the "RStudio Desktop Free", linked to https://posit.co/download/rstudio-desktop/. If you have not installed R itself earlier, install R first. Then you can download RStudio from the button of [DOWNLOAD RSTUDIO DESKTOP].
If your OS is Microsoft Windows, it links to RStudio-2025.05.0-496.exe (released on 5 May 2025) on 26 May 2025. After downloading, you can install RStudio automatically by clicking the icon and YES to query. In Microsoft Windows, starting icon will not be placed on the Desktop automatically, you may need to pin the RStudio icon after starting once.

If you make the shortcut to “C:/Program Files/Rstudio/bin/rstudio.exe” in “C:/Users/[your user name]/AppData/Roaming/Microsoft/Internet Explorer/Quick Launch” and show the quick launcher in the task bar.

Usually, when we use the RStudio, we set the specific project assigned to the folder. Choosing “New Project” from “File” menu and setting the project within the specific folder can set the default folder for that project. Within the folder, the project setting file with extension .Rproj is built. From next operation, by clicking that file (the file with extension .Rproj), RStudio will start, though when you run RStudio from start menu, desktop icon, or quick launcher, RStudio remember the previous setting.

Character codes are important issue for Japanese user, but RStudio can specify default character code for each project using “Tools>Project Options>Code Editing>Text Encoding” menu.

Installation and management of R packages

The excellent advantages of R include the enormous packages for new and/or special purpose analyses, freely available, developed by many specialists in the world. As https://r-pkgs.org/release.html says, the officially approved packages in the CRAN (https://cran.r-project.org/) are reliable but the developer has to overcome the difficulty to be accepted. The devtools package made the packages easily opened in GitHub (cf. https://kbroman.org/pkg_primer/pages/github.html). For such reasons, some packages are exclusively available at GitHub, but most famous packages are available at CRAN. If you want to install any specific package in CRAN, you need to enter the code **`install.packages("[known package name]", dep=TRUE)`** from R console and run.

For instance, if you want to install the Rcmdr package from CRAN, you need to enter below.

```
install.packages("Rcmdr", dep=TRUE)
```

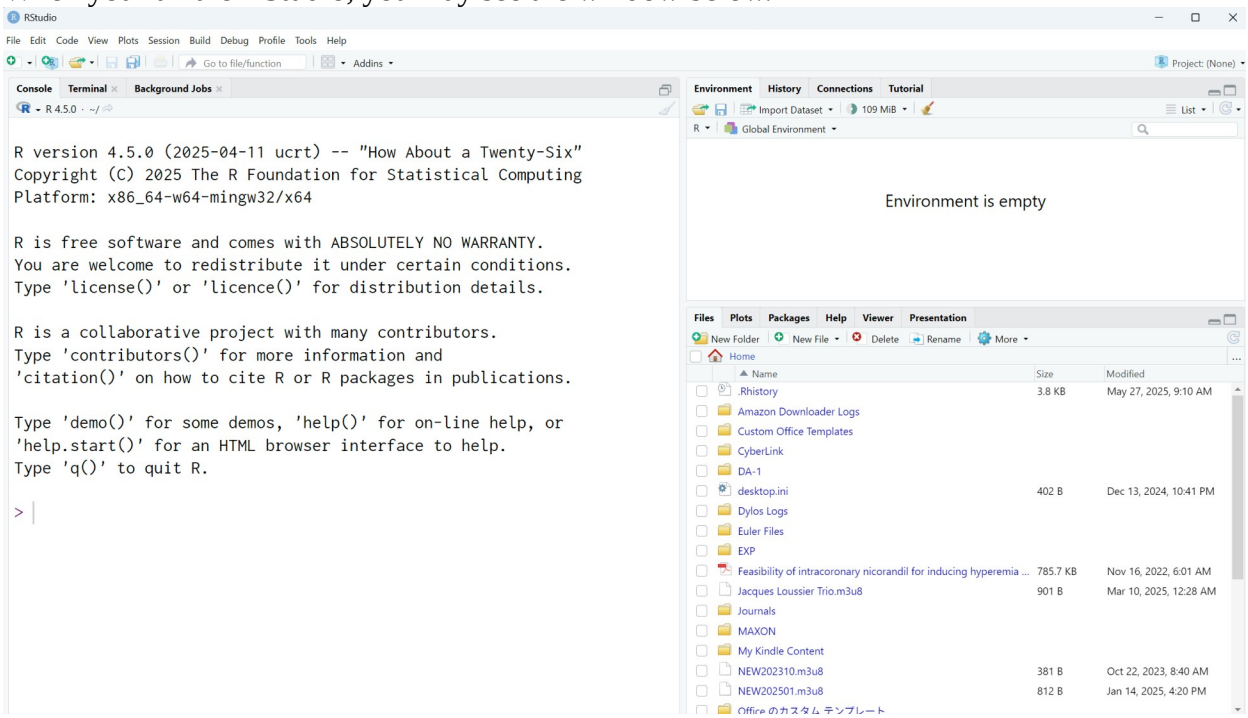
For the first time, you need to specify the name of mirror server. Usually cloud server or any server in your living country is recommended.

In the RStudio, from Tools menu, you can easily install and manage R packages (probably you can do it intuitively, therefore no explanation is necessary).

R basics

The CRAN provides various tutorial texts such as “R for beginners[1]” and “An introduction to R[2]”. Here the minimum explanation is given based on Microsoft Windows version with RStudio (the operation is almost the same in MacOS or Linux, but some details are different).

When you run the RStudio, you may see the window below.



The symbol “>” in the Console window is called as “prompt”. All interactive inputs to R are done via prompt, functions and statements followed by enter key (↵). If you fail to close parenthesis or wrongly press ↵ in the middle of functions or statements, the symbol of prompt changes to “+”, which means the one input line still continues. If you cannot exit from this “continuing line” status, press escape (Esc), then you can get back to new prompt (“>”).

If you directly use Rgui (not within RStudio), you can save the lines already inputted from the [File] > [Save History…], but in RStudio, this menu is inactive. Instead, in RStudio, you should make a script file first, usually via [File] > [New File] > [R Script] (or simply press [New file] icon below [File] menu, or press Ctrl Shift N at the same time), then you see the Script Editor window. You can enter any input lines to R as script here, then you can run the entered lines by selecting or put the cursor on the target line and click [→Run] icon (or press Ctrl ↵ at the same time). The script editor can edit, save and load the lines of R functions and statements (R scripts) at any time.

If you use the projects in RStudio, default folder to save and load all R scripts and data is the project’s folder. Of course, you can read and write scripts and data existing other folders or internet by specifying full paths or URIs, but I don’t recommend it except for very special situations.

---

1 https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf
2 https://cran.r-project.org/doc/manuals/R-intro.pdf

R objects

Basically we use R to analyze data in the way of (1) assigning the data to object by "**<-**", (2) applying the functions to draw graphs or to analyze to that object. Roughly considering, the object in R is a kind of variable. The name of objects should be composed of alphabet, 0-9, and period. Other characters such as underscore or 2 bytes characters are also applicable for the name of objects, but unusual.
The R is case-sensitive. X and x are treated as different objects. All names of functions and statements are composed of single byte characters, but 2 bytes characters can be used as the name of object. It causes very strange unexpected error. Instead of typing **hist(dat$HT)**, if you type ｈｉｓｔ（ｄａｔ＄ＨＴ）, error message will be raised. If you type hist(dat$HT, xlab="height (cm)"), ", " is double byte character and thus nothing happen (even error message will not appear). This kind of error is very difficult to detect. By typing **ls()**, you can get the list of all objects already defined in that session. There are several types in R objects, as shown below.

Various scalar types

NULL: The value showing nonexistence. Different from NA (missing).

Logical: TRUE or FALSE. Integer number 1 and 0 can be transformed to logical using as.logical(). Equations returns logical value. "**5<4**" is FALSE. "**1+2==3**" is TRUE.

Integer number: For example, -1564, 0, 1, 45671, and so on. Too large integer number is automatically converted to real number (floating point). To convert other type scalar to integer, **as.integer()** can be used. **as.integer(TRUE)** is 1. **as.integer("455")** is 455. If you connect 2 integer numbers with "**:**", a vector of integer numbers is generated. **2:4** means **c(2, 3, 4)**.

Numeric (real number): 0.1, pi, sqrt(2), 1e+7, and so on. The "**pi**" is the built-in constant of 3.14159265. The **sqrt()** is the function to calculate square-root. The "e+" means 10^{th} powers. 1e+7 means 10000000. If you want to convert other type to numeric, as.numeric() can be used.

Complex number: Not used in statistics, but 1+1i, 0+0i, or any other complex numbers can be treated in R. To convert other types to complex, **as.complex()** can be used.

Characters: By quoting anything with double quote or single quote symbols, it can be treated as character. "abc", "1", and so on. **LETTERS** means upper case alphabets. Therefore, **LETTERS[24:26]** means **c("X", "Y", "Z")**. Any objects can be converted to characters by as.character().

Factors: It doesn't make sense as scalar, because the factor type is composed of integer levels with the names (labels) of each level. This type is applicable to use categorical variable. Defining it by **factor()** or converting other types with **as.factor()**. Please read the explanation below.

To understand the factor type, let's consider an example. ABO blood type is composed of 4 categories of "A", "B", "O", and "AB". Therefore, the variable for ABO blood type is a factor with 4 levels, each blood type is considered as the name of levels. If you read it from tab-delimited text file, you can use the function **read.delim()** with the option "**stringsAsFactors = TRUE**", when the levels

are automatically assigned due to alphabetical order. In this case, 1st level is “A”, 2nd level is “AB”, 3rd level is “B”, and 4th level is “O”. If you don’t use that option, the data is read as characters.
If you want to change the levels, you can use **`factor()`** function as shown below.

```
set.seed(123) # initialize random number seed
bloodtypeC <- sample(c("A","O","B","AB"), 100, rep=TRUE)
bloodtype <- factor(bloodtypeC, levels=c("A", "O", "B", "AB"))
bloodtype <- factor(sample(1:4, 100, rep=TRUE),
 labels=c("A", "O", "B", "AB")) # This is better way to specify order
```

When the **`as.integer()`** function is applied to factor type object, integer vectors showing the levels of factors are returned. If the levels have apparent order, we can use ordered factor type by **`ordered()`** or **`as.ordered()`**. Sometimes we need to categorize the continuous numbers such as age or height into ordered categories. If we have the height as numeric variable, we sometimes need to categorize by every 5cm. Let’s generate hypothetical heights data for 20 people whose average is 160 cm with 5 cm of standard deviation. Then categorize that into factor object for every 5cm from 150 to 180 cm using **`cut()`** function.

```
set.seed(123)
height <- round(rnorm(20, 160, 5), 1)
hc <- cut(height, seq(150, 180, by=5))
print(data.frame(height, hc))
```

The function **`cut(continuous_variable, interval_vector)`** converts any continuous variable into a categorical variable according to the interval vector. The **`seq(min_value, max_value, by=interval_length)`** function, used to define the interval vector, generates an interval vector by dividing the range from the minimum value to the maximum value by the interval length. If you use **`len=number_of_intervals`** instead of **`by=`**, it will divide the range from the minimum value to the maximum value into the specified number of equal intervals. hc becomes a factor-type variable where the minimum interval is (150, 155] and the maximum interval is (175, 180]. By default, the boundaries of the intervals are "greater than ~ and less than or equal to ~". If you want to use the Japanese style of "greater than or equal to ~ and less than ~", specify the option **`right=FALSE`** within the **`cut()`** function. Furthermore, if you use **`hco <- ordered(hc),`** it will become an ordered factor type.

The R objects can have various structures such as vector, matrix, table, list, and data.frame. Vector, matrix and table can include the single type object as its elements, but list and data.frame can include several different types of objects. In addition, the objects in S4 and S5 classes can have the slots. The map data objects is one of the S4 classes, but it cannot be explained in this text because the author doesn’t understand well.

Vector: It’s a set of scalars. Including all elements of a vector within **`c()`** separated by comma(**`,`**). If the included elements have different types, the types are automatically converted. For example, **`c(NULL, FALSE, 11, 8.23, 5+2i, "statistics")`** is automatically converted to **`c("FALSE", "11", "8.23", "5+2i", "statistics")`**. However, **`NULL`** is omitted from the set, because it means empty. To extract the elements from a vector can be done using brackets, **`[]`**. When **`x`** is **`2:4`**, **`x[3]`** returns **`4`**.

Matrix: A vector with dimension is a matrix. **`matrix(X, NROW, NCOL)`** can convert a vector **`X`** to a 2 dimensional matrix with **`NROW`** rows and **`NCOL`** columns. When **`NCOL`** is not given, **`length(X)/NROW`** is assumed. The alignment of a vector is from top left to bottom left, then 2nd column's top, go down, and so on. When the option of **`byrow=TRUE`** is given, the alignment is from top left to top right, then move to 2nd row. The NROW and NCOL must be the positive integer numbers. **`matrix(X, 1)`** looks similar to a vector **`X`** itself, but different because a matrix has dimension. Reference to any element of a matrix can be done using brackets with comma **`[the order in row, the order in column]`**. eg., **`matrix(1:9, 3, 3)[3, 1]`** is **`3`**, **`matrix(1:9, 3, 3)[1, 3]`** is **`7`**. However, when matrix is compared with other objects (including not only matrix but also vector or scholar), each element of matrix is compared. For example, **`1:4 == matrix(1:4, 1, 4)`** returns the matrix with 1 row and 4 columns, **`TRUE TRUE TRUE TRUE`**. For matrix with 3 dimensions, **`array(X, dim=c(NROW, NCOL, NSTRATA))`** can be used. X is a vector. Elements of X are aligned from the 1st strata's top-left to going down, then next columns top, going down, … until the bottom-right, then moving to the 2nd strata. Such 3D matrix is useful for some special functions such as **`mantelhaen.test()`**, which calculates Cochran-Mantel-Haenszel's pooled odds ratio and conducts Cochran-Mantel-Haenszel's pooled chi-square test across strata. If you have 2 or more vectors with same length, using **`cbind()`** or **`rbind()`**, you can get matrix. In addition, matrices can be transposed by **`t()`**.

Table: It's a kind of matrix, but has attributes of "table". Basically all elements of a table are integer number. Rows and columns have names. Usually, table object is generated as the result of **`table()`** or **`xtabs()`** to make contingency tables, but if you apply **`attr()`** to matrix type objects such as **`attr(X, "class") <- "table"`**. R is object-oriented language so that some generic functions in R can change the default behavior by the type of objects given for that function. For example, when an object is given to **`plot()`** function, usually scattergram is drawn, but if the object type was table, **`mosaicplot()`** is automatically applied and mosaic plot is drawn.

List: A list type object can include any types, any numbers objects within it, by the function **`list()`**. A list can include even other lists, which can be nested. Each list item can be named. For example,
`X <- list(A=1:3, B=c("X", "Y"), C=TRUE)`
can generate a list X including 3 list items, A of integer vector, B of character vector and C of logical scalar as those are. Referring the list item can be done by $["the name of list item"] or [["the name of list item"]] or [[the order of list item]]. In this example, **`X$A`** returns c(1, 2, 3), **`X[["B"]]`** returns c("X", "Y"), and **`X[[3]]`** returns TRUE. If you need to refer the element within the list item, the brackets can be used again. For exmaple, both **`X$B[2]`** and **`X[[2]][2]`** indicates the 2nd element of list B, that is "Y".

Data.frame: The object with type **`data.frame`** is a special kind of **`list`**, where all list items are vectors with the same length. It looks like a table, but can include several different types of vectors, which distinguish data.frame from matrix or table. Referencing a part of data.frame is possible by 2 ways, the way used in matrix (such as **`X[i, j]`**) and the way used in list (such as **X$A[i]**). If you read the data from external tab-delimited text file using **`read.delim()`** or comma-separated text (csv) using **`read.csv()`**, the resulted object becomes **`data.frame`**. The object **`X`** of matrix type can be converted to data.frame type by **`as.data.frame(X)`**. The function **`subset(X, expr)`** can extract a part of data.frame, where only the rows with **`expr`** being **`TRUE`** are extracted as a data.frame.

To get information of objects, the following functions are available.

Mode: The **`mode(x)`** inspects the type of object **`x`** if **`x`** is scalar, vector or matrix. If **`x`** is list or data.frame, “list” will be returned.

Structure: The **`structure(x)`** or its abbreviation **`str(x)`**, returns the data structure information of list or data.frame object **`x`**, such as type and length of objects included in **`x`**.

Attributes: The **`attributes(x)`** or its abbreviation **`attr(x)`** inspects the class of object **`x`**.

Length: The **`length(X)`** returns the length of the object, which means the number of elements for vector, number of the list items for list, and the number of variables for data.frame. If you need to get the length of character strings **`S`**, **`nchar(S)`** is available. For example, length(“happy”) returns 1, but nchar(“happy”) returns 5.

Names: \item[names] \verb!names()!\index{names()}は、\verb!rownames()!、\verb!colnames()!、\verb!dimnames()!とともに、スカラーに名前を付けたり、ベクトルやリストや行列やテーブルやデータフレームに含まれる変数名を参照したり、それらを付値によって改変する目的で用いる。オブジェクト\verb!x!の値が 1 だとして、この\verb!x!に、例えば\verb!"test"!という名前を付けるには、\verb!names(x) <- "test"!とする。ベクトルの場合、例えば、リンゴが 5 個、ミカンが 3 個、メロンが 2 個、葡萄が 10 個あることを表現したければ、\verb!x <- c(5, 3, 2, 10)!としてから\verb!names(x) <- c("apple","orange","melon","grape")!とすればよい。名前を付ける利点は、それによる参照ができるようになることで、この場合なら、メロンの個数を知りたいとき、\verb!x["melon"]!という形で参照できる。行列またはデータフレーム\verb!X!について、\verb!rownames(X)!で\verb!X!の行の名前を参照できるし、\verb!rownames(X) <- c("A", "B", ...)!のようにすれば行名を付けることができる。\verb!colnames(X)!で列名が参照でき、\verb!colnames(X) <- c("X", "Y", ...)!で列名を付けることができる。

Basic Grammar in R

Inputs until carriage-return or “;” to prompt is treated as single function or statement. The most basic functions and statements are summarized below.

\item[終了] \verb!q()!\index{q()}
\item[付値] \verb!<-!\index{<-@$<-$}\par
例えば、1、4、6 という 3 つの数値からなるベクトルを\verb!X!というオブジェクト（変数）に保存するには次のようにする。
\begin{screen}\small\begin{verbatim}
X <- c(1, 4, 6)
\end{verbatim}\end{screen}
\item[注釈] \verb!#!\index{#@\#}より後は行の終わりまで注釈となり実行されない
\item[区切り] \verb!;!は改行の代わりになり、1 行の中に 2 つ以上の関数や文を書ける
\item[ブロック] \verb!{!から\verb!}!まではブロック\index{ぶろっく@ブロック}となり、間に改行があっても 1 つの塊として扱われる
\item[関数の適用] 関数にオブジェクト（とオプション）を与えると結果が返ってくる。例えば、上の\verb!X!に対して、合計を計算する関数\verb!sum()!\index{sum()}を適用するには、\verb!sum(X)!とす

れば、11 という結果が返ってくる。関数は入れ子にできるし、関数の結果をオブジェクトに付値することもできる。
\item[定義] \verb!function()!\index{function()}\par
複雑な計算を 1 つの関数\index{かんすう@関数}として自分で定義\index{ていぎ@定義}することができる。関数定義は何行にも渡って行うことができ、最終行の値が戻り値となる。関数内の変数は局所化されているので、関数内で変数に付値しても関数外には影響しない。関数内で変数の値を本当に変えてしまいたいときは、通常の付値でなくて、\verb!<<-!（永続付値）\index{えいぞくふち@永続付値}を用いる。例えば、ベクトル X の平均\index{へいきん@平均}と標準偏差\index{ひょうじゅんへんさ@標準偏差}を計算する関数\verb!meansd()!の定義は次の通り。
\begin{screen}\small\begin{verbatim}
meansd <- function(X) { list(mean(X), sd(X)) }
\end{verbatim}\end{screen}
\item[ヘルプ] \verb!?!\index{?@$?$}\par
例えば、t検定の関数\verb!t.test()!\index{t.test()}の解説をみるには、\verb!?t.test!とする。見出し語が不明で説明文中に出てくる単語を検索したいときは$??$\index{??@$??$}を使う。例えば、コクラン＝マンテル＝ヘンツェルの要約カイ二乗検定をする関数名を忘れてしまったときに、\verb!??Cochran!と打てば、\verb!mantelhaen.test()!という関数名が見つかる。
\item[使用例] \verb!example()!\par
多くの関数はヘルプに用例が含まれていて、\verb!example()!\index{example()}で実行することができる。例えば\verb!example(lm)!とすれば、線形回帰分析\index{せんけいかいきぶんせき@線形回帰分析}の関数\verb!lm()!\index{lm()}の使用例が表示される
\end{description}

R は S 言語\index{S げんご@S 言語}のサブセットの実装と言われている通り、S 言語の文法でループ\index{るーぷ@ループ}や条件分岐\index{じょうけんぶんき@条件分岐}などの制御構造\index{せいぎょこうぞう@制御構造}が書ける。簡単に説明しておく。

\begin{description}
\item[ループ] \verb!for () { }!\index{for}によるのが普通である。例えば、
\begin{screen}\begin{verbatim}
T <- 0
for (i in 1:3) {
 T <- T+i
}
\end{verbatim}\end{screen}とすると、最初は 0 である T に 1、2、3 が順に足されて 6 に変わる。条件分岐によりループを途中で抜けたいときは\verb!break!を使う。ループの終了条件を予め決められない場合は、\verb!while( ) { }!を使うことができる。
\item[条件分岐] \verb!if () { } else { }!\index{if}が条件分岐の基本形である。\verb!( )!内に入れる条件文には次のようなものがある。条件文がベクトルならば最初の要素のみ使われることに注意。
\begin{screen}\begin{verbatim}
if (A==B) {} # A と B が等しいとき{}内を実行
if (A>B) {} # A が B より大きいとき{}内を実行
if (A>=B) {} # A が B より大きいか等しいとき{}内を実行
if (A<B) {} # A が B より小さいとき{}内を実行
if (A<=B) {} # A が B より小さいか等しいとき{}内を実行
if (x %in% A) {} # x が A の要素に含まれれば{}内を実行

\end{verbatim}\end{screen}
ベクトルの各要素に対して条件判定させ、新たなベクトルを作るには\verb!ifelse()!\index{ifelse()}を用いることができる。例えば、\verb!x <- c(1, 1, 2, 1, 2, 2, 1)!であるとき、1 ならば"M"、2 ならば"F"に置き換えた新しい変数\verb!y!を作りたいときは、\verb!y <- ifelse(x==1, "M", "F")!とする\footnote{もっとも、この場合は \begin{screen}\begin{verbatim}y <- factor(x, labels=c("M", "F"))\end{verbatim}\end{screen} とファクター化する方が普通である。}。カテゴリ変数の再カテゴリ化\index{さいかてごりか@再カテゴリ化}も、\verb!ifelse!と\verb!%in%!\index{\%in\%}を使うとやりやすい。
\end{description}

Perspectives of statistical analyses by the type of data

The data obtained by questionnaire survey

Data obtained from questionnaire surveys is fundamentally categorical data. It is necessary to clarify whether one is asking about knowledge, attributes, behavior, or perception. To ascertain knowledge, a test with clearly defined correct and incorrect answers should be administered. Test scores can be treated as continuous variables if they approximate a normal distribution. To investigate attributes and behavior, questions must be devised that can ask about facts as unambiguously as possible. Although continuous data may be obtained in some cases, such as age or sleep duration, attributes and behavior mostly result in categorical data.

For perceptions, a Likert scale[3] is often used. Whether to use a 3-point or 5-point Likert scale, for instance, depends on the objective. A 3-point scale (e.g., options like {1. Disagree, 2. Neither agree nor disagree, 3. Agree}) or a 5-point scale (e.g., {1. Strongly disagree, 2. Somewhat disagree, 3. Neither agree nor disagree, 4. Somewhat agree, 5. Strongly agree}) are commonly employed. These are themselves analyzed as ordinal scales. However, when the total score of multiple similar question items is used as some kind of scale, it is often treated as a continuous variable. In such cases, it is necessary to confirm whether these question items indicate a single latent factor using methods such as Cronbach's alpha coefficient. If the alpha coefficient is not generally 0.7 or higher, the reliability as a single latent factor is considered low. In some cases, it may be necessary to perform factor analysis and re-examine the latent factor structure.

Notes in setting question items

- The question items must include needed and enough items, all of which have to relate with research purpose or working hypothesis.
- Generally, the process involves first breaking down the main subject of the survey objective into several dimensions. Then, each of these is further decomposed into several more detailed dimensions. This procedure is repeated, and the finally subdivided elements become the survey items, which are incorporated into the questionnaire as headings or questions.
- There's a tendency to want to include anything that seems interesting or valuable, but careful consideration is necessary:
  - Will data be obtainable from that item?
  - Even if data is obtained, can it be used for analysis?[4]
  - Is its importance not low considering the overall balance?
  - Will it not cause resistance, aversion, or bewilderment in the respondents?
- The principle is to include the bare essentials, plus introductory questions for sensitive topics, and a limited number of seemingly redundant questions to confirm responses to other questions.

---

3 Likert scale: Respondents choose from several options to indicate the extent to which they agree with (or find applicable) a presented statement, and this choice is scored.

4 To confirm this point, dummy tables (provisional tables created to visualize how the data would be tabulated and summarized if obtained) are useful. They should be created if possible.

Notes in wording

- Simply and easy to be correctly understood by the participants. eg. When you need to get the information related to age, asking date of birth rather than age is better.
- Pay attention to the words used in daily conversation, adverbs, pronouns. eg. “What kind of cloths do you wear?” cannot clarify the contents of “kind”. “Why is it?” may cause misunderstanding in what “it” specifies.
- Pay attention to the following points.
    - Be careful with common nouns and proper nouns. For example: "How many copies of the newspaper do you subscribe to?" doesn't specify the type of newspaper, and it's ambiguous whether asking about the number of copies includes those for customer service in a business.
    - Pay attention to differences in imagery due to the respondent's social class or environment. For example: Even if you say "bath," a student living in a boarding house might imagine a public bathhouse.
    - Avoid difficult terms or technical jargon. If you must use them, provide clear definitions. For example: It's reported that in a survey conducted in the US, when asked about the completely fictitious "Metallic Metals Act," 70% of respondents answered that "it should be investigated by the federal or state government." Respondents are reluctant to say they "don't understand the words."
    - Avoid stereotypical words. For example: The image conjured by the word "leftist" can vary greatly among respondents.

Notes in statements

- Avoid using excessive adjectives.
- Devise sentences while considering that respondents tend to affirm questions ("yes" tendency).
- Be especially careful that introductory clauses in complex sentences do not become leading (e.g., prestige suggestion effect). For example: A question like "It is said in the world that ○×, but you..." distorts the answer by borrowing the prestige of public opinion.
- Avoid questions that are limited by units. For example: With the question "About how many books do you read in a month?", people who read 2-3 books a year are likely to be forced into categorizing themselves as 0 or 1.
- Questions whose sentence content includes two or more points (Double-barreled questions[5]) should be broken down into a group of questions, one for each point.
- Do not ask questions based on detailed past memories.
- Avoid questions with negative phrasing as they are ambiguous. For example: In the question "Should the sale of the municipal zoo to a canning company be stopped, or do you think it should not?", the answer "I think it should not" is ambiguous as to whether it means "it should not be sold" or "it should not be stopped."
- Do not ask overly bizarre questions. For example: A sudden question like "If you were to live on Mars..." could undermine the reliability of the entire survey.

5 The term barrel usually means a cask, but it's different in this case. According to Ishikawa, Sato, and Yamada (1998) "The Power to See the Unseen [Mienai-Mono-Wo-Miru-Chikara]" Yachiyo Shuppan, p.284, "Incidentally, double-barrel refers to a double-barreled gun, which is designed to fire two bullets at once."

Types of questions

- Is the question personal or general?
- Are you asking about awareness or actual conditions?
- Are you asking for an opinion or testing knowledge? Questions that test knowledge can be used as filter questions to only ask for opinions from those who possess the relevant knowledge. However, if the primary purpose is to test knowledge, a test should be administered.
- Are you asking about normal behavior or behavior on a specific date/time? For example, when conducting a dietary survey, the results usually differ between the 24-hour recall and the food frequency questionnaire (FFQ).
- Are you asking with a single question or capturing information with a set of questions? To inquire about the constructs that cannot be grasped with a single question, it's typical to use a validated set of questions and use their total score as the score for the construct (Note: Even if the set of questions has already been validated, internal consistency must be checked for the data obtained using Cronbach's alpha).
- For those who answered yes or no to a specific question, a second, leading question is asked to make them overturn their judgment, thereby measuring the strength of the yes or no response to the first question. This second question is called as biased question. For example: If someone answers "yes" to "Are you going to vote in the upcoming general election, or not?", you might ask, "What if it rains on election day / What if you have something else to do on election day?". If someone answers "no", you might ask, "What if an acquaintance invites you?". This is difficult due to the significant impact of wording, but if done well, it can provide a sharper assessment than scoring with a Likert scale from the outset.

Types of answering

- Free answer: Easy to ask, but analysis is sometimes difficult.
- Precoded free answer: Getting answers in free-format, but the researcher prepares the set of expected answers as precoded categories, and the enumerator checks one of the those categories based on answers given by respondents.
- Choosing answer: The researcher have to prepare candidate answers, then respondents choose answers from those candidates. Yes/No = Binary, Rating, Qualitative/quantitative choice from 3 or more candidates
- Ranking
- Multiple choice: At the time of data entry, usually the question with multiple choice is separated into each candidate items with Yes/No.

The indicators to show internal consistency as reliability of the scale

There are many kinds of scales, but in the studies in health and medicine, Likert scale is most commonly used. Cronbach's alpha is the most famous one, but recently McDonald's omega is also used. However, there are many approaches to internal consistency.

Scales created by Item Analysis: This item selection method adopts only items with high correlation to the quantitative characteristic being measured. Typically, good-poor analysis is performed. The procedure involves, in a pre-test, assigning tentative scores to all candidate items for inclusion in the scale and summing them. Then, items where there was a difference in response categories between the upper group (total score greater than the third quartile) and the lower group (total score smaller than the

first quartile) are adopted as constituent questions for the scale. (Usually, response categories are combined into a dichotomous format, and a difference is considered to exist if the proportion of people who responded positively in the upper group is statistically significantly higher than in the lower group. If there are many response categories and it's difficult to convert to a dichotomous format, a t-test for mean differences or a Wilcoxon rank-sum test may be used to check for significance.) For each adopted item, the scores assigned to each response category can be determined by: (1) arbitrary assignment, (2) requesting a group of judges, (3) Likert's sigma method (assuming a normal distribution and assigning scores as $z_i = (y_{i-1} - y_i)/(p_i - p_{i-1})$. Likert showed that if simple total scores are used, this is almost equivalent to assigning integers sequentially from 1 to each category), (4) Seavert et al.'s sigma method (different formula but also assumes a normal distribution, so scores are almost identical to Likert's), or (5) Guilford's method.

Scales created by Scale Analysis: A method devised by Guttman, using a scalogram.

Scales created by Factor Analysis: Basically, scales are constructed by summing (or weighted summing) items that are classified into the same factor (have large factor loadings on the same factor). The reliability of this scale is usually examined using the alpha coefficient.
In this text, I will briefly explain only **Cronbach's** α and **McDonald's** ω.

When attempting to measure some concept using a questionnaire, many concepts cannot be directly elicited. Therefore, we try to combine multiple questions to grasp individual differences more precisely. For example, if you want to gauge a person's affinity for nature:

(1) Do you like or dislike nature? (Like, Somewhat like, Somewhat dislike, Dislike)

This question alone would only divide respondents into four groups (if quantified as an ordinal scale, "Like" could be 4 points, "Dislike" 1 point, resulting in four levels from 1 to 4 points). However, if you add:

(2) On holidays, do you prefer spending time at the sea or in the mountains, or at a movie theater or amusement park? (Sea/Mountains, Somewhat Sea/Mountains, Somewhat Movie Theater/Amusement Park, Movie Theater/Amusement Park)

And treat this as an ordinal scale where "Sea/Mountains" is 4 points and "Movie Theater/Amusement Park" is 1 point, then calculating the total score for answers to (1) and (2) could classify respondents into seven groups, ranging from 2 to 8 points, allowing for a more detailed understanding. Furthermore, adding:

(3) Are you attracted to a job observing wildlife in an uninhabited jungle, or not? (Attracted, Somewhat attracted, Somewhat not attracted, Not attracted)

This would increase the range to 10 levels, from 3 to 12 points. If we consider this total score as a scale representing "affinity for nature," it is expected that the three items, being sub-concepts constituting the same concept, will elicit responses with similar tendencies.
That is, someone who answered "Like" in (1) would likely answer "Sea/Mountains" in (2) and "Attracted" rather than "Not attracted" in (3). If consistent response tendencies are obtained for questions constituting the same concept, the scale represented by their total score is considered to have high **reliability**.

One indicator for examining the relationship between multiple variables (items) is the Pearson's correlation coefficient. The correlation coefficient r_{xy} between variable x and variable y is defined as follows, where x_i is the i-th person's response to x, y_i is their response to y, $\overline{x}$ is the mean response for x, $\overline{y}$ is the mean response for y, and n is the total number of respondents:

$$r_{xy} = \frac{s_{xy}^2}{(s_x s_y)} \quad \text{where} \quad s_x = \sqrt{\frac{\sum_i (x_i - \overline{x})^2}{n-1}} \quad s_y = \sqrt{\frac{\sum_i (y_i - \overline{y})^2}{n-1}} \quad s_{xy}^2 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

The correlation coefficient ranges from -1 to 1, being 0 when there is no relationship, and 1 when (x_i, y_i) plots perfectly on a straight line with a positive slope in the xy-plane.

One method for checking whether consistent answers were obtained for the three questions above is the split-half method. For example, by dividing all questions constituting the same concept into two, such as the total score for questions (1) and (2) as variable x_{12} and the score for question (3) as variable x_3. If the correlation coefficient between x_{12} and x_3 is $\mathrm{r}_{x_{12}x_3}$, then the reliability coefficient $\alpha_{x_{12}x_3}$ for these questions is given by Spearman-Brown's formula (typically, items are split into odd-numbered and even-numbered items).

$$\alpha_{x_{12}x_3} = \frac{2r_{x_{12}x_3}}{1+r_{x_{12}x_3}}$$

However, there are other ways to split the items, such as the score for (1) and the total score for (2) and (3). If there are three or more sub-concepts, and you want to use Spearman-Brown's formula to determine if there's a consistent tendency in these responses, multiple α values would be calculated (as many as there are combinations of splitting n items into two groups). In this case, $\alpha_{x_1x_{23}}$ and $\alpha_{x_{13}x_2}$ would also need to be calculated. Then, Cronbach's α aims to summarize all these possibilities. If the total score of (1), (2), and (3) is taken as variable x_t representing "affinity for nature," and the scores for (1), (2), and (3) are variables x_1, x_2, and x_3 respectively, then Cronbach's α is calculated as:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} s_{x_i}^2}{s_{x_t}^2}\right)$$

where k is the number of items. In our example with 3 items, it's:

$$\alpha = \frac{3}{3-1}\left(1 - \frac{s_{x_1}^2 + s_{x_2}^2 + s_{x_3}^2}{s_{x_t}^2}\right)$$

It's the item count divided by (item count minus one), multiplied by one minus the sum of the variances of the scores for each item divided by the variance of the total score. A Cronbach's α coefficient of 0.8 or higher is generally considered to indicate sufficient, and 0.7 as fair, internal consistency (reliability) for that set of items. Notably, Cronbach's α coefficient yields the same value as the average of the α values calculated using Spearman-Brown's formula for all possible combinations.

Suppose the results of applying the above questions to 10 people were obtained as shown at https://minato.sip21c.org/advanced-statistics/cronbach.txt. If this tab-delimited text file is loaded into a data frame **x**, then after loading the **fmsb** package by **library(fmsb)**, **CronbachAlpha(x)** will show that Cronbach's α coefficient is 0.8027. If you load the **psych** package and use **alpha(x)**, you will obtain not only the point estimate but also various other values, such as the 95% confidence interval.

In addition, by using the **`reliability()`** function included in the **`semTools`** package (which you would need to install), you can obtain not only Cronbach's α but also five other types of reliability coefficients: ω_1, ω_2, ω_3, and AER. McDonald's omega (ω) is considered an alternative or improvement over Cronbach's alpha[6], especially when the assumptions of alpha (like tau-equivalence, meaning all items measure the same latent construct with equal strength) are violated. ω is particularly useful in **factor analysis** and **structural equation modeling** as it accounts for the actual factor structure and item loadings, providing a more accurate estimate of composite reliability when items contribute unequally to the overall scale score. ω_1 (sometimes denoted ω_h for hierarchical omega) and ω_2 (often denoted ω_t for total omega) are common forms, with ω_3 also appearing in some contexts.

Typical flows and layouts of questionnaire

\begin{itemize}
\item 質問の順序の原則\par
\begin{enumerate}
\item 答えやすい質問は前
\item 関連する事柄や似ているものは集める（システマティックにつくる）。ただし、それゆえにキャリー・オーバー効果\index{きゃりーおーばーこうか@キャリー・オーバー効果}（回答が、それまでの質問項目の影響を受けてしまうこと）が問題となる場合もある。
\item 対象者を限定する枝分かれ質問（サブクエスチョン\index{さぶくえすちょん@サブクエスチョン}）で間違いにくい順番を工夫する
\end{enumerate}
\item タイトル：反発を起こすものは避ける
\item 調査主体や連絡先の明記。
\item 挨拶
\item 記入上の注意
\item 調査票についての処理の記録欄：コーディングで使う
\item 小見出しや説明：対象者に調査の順序をわかってもらうための説明
\item 質問番号：論理的階層性が明確な方がよい
\item 回答上の指示：【】に入れるとか書体を変えるなど、質問との区別がはっきりするように。
\item お礼の挨拶
\item 調査員判定
\item 最終レイアウトとページ数：最後のページがだいたい一杯におさまるようなレイアウトにし、通しのページ番号を振る。
\end{itemize}

The perspectives to analyze the data from questionnaire survey

6 Hayes AF, Coutts JJ (2020) Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. *Communication Methods and Measures*, 14(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

カテゴリ変数同士の関係をみることが多いので、多様なクロス集計\index{くろすしゅうけい@クロス集計}をする必要がある（R では\verb!table()!\index{table()}や\verb!xtabs()!\index{xtabs()}で可能）。クロス集計\index{くろすしゅうけい@クロス集計}においては、変数間の独立性\index{どくりつせい@独立性}の検定をフィッシャーの直接確率\index{ふぃっしゃーのちょくせつかくりつ@フィッシャーの直接確率}（R では\verb!fisher.test()!\index{fisher.test()}を使う）で行うことが多い。関連の強さ\index{かんれんのつよさ@関連の強さ}は、オッズ比\index{おっずひ@オッズ比}（\verb!fmsb!\index{fmsb}パッケージの\verb!oddsratio()!\index{oddsratio()---fmsb}関数や\verb!vcd!\index{vcd}パッケージの\verb!oddsratio()!\index{oddsratio()---vcd}関数などを使う）や四分相関係数\index{しぶんそうかんけいすう@四分相関係数}（R では\verb!vcd!\index{vcd}パッケージの\verb!assocstats()!\index{assocstats()---vcd}関数や、\verb!polycor!\index{polycor}パッケージの\verb!polychor()!\index{polychor()---polycor}関数で計算できる）などで評価することが多い。３つ以上のカテゴリ変数間の関係を見るときは、コクラン＝マンテル＝ヘンツェルの要約カイ二乗検定\index{こくらんまんてるへんつぇるのようやくかいにじょうけんてい@コクラン＝マンテル＝ヘンツェルの要約カイ二乗検定}や（\verb!mantelhaen.test()!\index{mantelhaen.test()}を使う）、ロジスティック回帰分析\index{ろじすてぃっくかいきぶんせき@ロジスティック回帰分析}（\verb!glm()!\index{glm()}で計算できる。適合度の指標としては\verb!fmsb!\index{fmsb}パッケージに入っている\verb!NagelkerkeR2()!\index{NagelkerkeR2()---fmsb}により Nagelkerke のR^2や\verb!AIC()!\index{AIC()}を計算する）を実行する。

リッカート尺度による聞き取り結果をスコア化する場合は、クロンバックのα係数\index{くろんばっくのあるふぁけいすう@クロンバックのα係数}（\verb!fmsb!\index{fmsb}パッケージの\verb!CronbachAlpha()!\index{CronbachAlpha()---fmsb}関数や、\verb!psych!\index{psych}パッケージの\verb!alpha()!\index{alpha()---psych}関数で計算できる）が小さければ因子分析（詳細は後述）する場合もある。スコアは量的変数として扱うので、カテゴリ間にスコアの差があるかどうかを調べるには、t検定\index{t けんてい@t検定}（\verb!t.test()!\index{t.test()}を用いる）や一元配置分散分析\index{いちげんはいちぶんさんぶんせき@一元配置分散分析}（\verb!aov()!\index{aov()}や\verb!oneway.test()!\index{oneway.test()}を用いる）を行うこともある。

最終的には構造方程式モデル\index{こうぞうほうていしきもでる@構造方程式モデル}（詳細は後述）を当てはめる場合も多い。

なお、質問紙への回答の信頼性\index{しんらいせい@信頼性}を確かめるために同じ質問紙調査を対象者 1 人につき 2 回ずつ実施する（あるいは、同じ対象者への異なる評価者による評点\index{ひょうてん@評点}があるとき、各項目について 2 人分の評点が付される）ことがある。この場合も、回答がカテゴリであれば、2 つの別々の質問項目の場合と同じ形でクロス集計表\index{くろすしゅうけいひょう@クロス集計表}を作ることができるが、独立でないのは当然なので、フィッシャーの直接確率\index{ふぃっしゃーのちょくせつかくりつ@フィッシャーの直接確率}などを計算しても意味は無い。むしろ、偶然では考えられないほど一致しているかという、一致度\index{いっちど@一致度}を計算すべきである。典型的な一致度の指標は Cohen のκ係数\index{かっぱけいすう@κ係数}である。完全一致の場合 1、偶然と同じ一致度で0、完全不一致で-1となる。\verb!fmsb!\index{fmsb}パッケージの\verb!Kappa.test()!\index{Kappa.test()---fmsb}関数に行数と列数が等しいクロス集計表オブジェクトを与えると自動的に計算され、一致度がどの程度かという目安も表示される。

これと同じように、同じ質問が繰り返される場合であっても、2 回の調査の間に何らかの介入があって、介入効果\index{かいにゅうこうか@介入効果}によって回答が偶然では考えられないほど変化したかを知りたい場合も、フィッシャーの直接確率のような独立性の検定は使えない。代わりに用いるのはマクネ

マーの検定\index{まくねまーのけんてい@マクネマーの検定}であり、\verb!mcnemar.test()!\index{mcnemar.test()}関数に行数と列数が等しいクロス集計表オブジェクトを与えれば実行できる。

The data from experimental (mostly in laboratory) study

治験\index{ちけん@治験}を含む実験\index{じっけん@実験}の場合、カテゴリデータは曝露\index{ばくろ@曝露}の有無など所与の条件であることが多い。毒性試験\index{どくせいしけん@毒性試験}では毒性発現の有無、疾病発生の有無、死亡か生存かといった２値データをアウトカム\index{あうとかむ@アウトカム}として用いる場合もある。それらを除けば、実験で得られるデータは、概ね数値型の測定値である。測定限界\index{そくていげんかい@測定限界}と有効数字\index{ゆうこうすうじ@有効数字}に注意する必要がある。

実験では、統計解析方法は実施前に決めておくのが原則である。新薬の有効性であれば分散分析\index{ぶんさんぶんせき@分散分析}、毒性試験ならばアウトカム発生までの時間に対する生存時間解析\index{せいぞんじかんかいせき@生存時間解析}、あるいは用量反応関係についてのプロビット解析\index{ぷろびっとかいせき@プロビット解析}またはロジット解析\index{ろじっとかいせき@ロジット解析}による LD50 や ED50 の推定、アウトカムとして量的な効果をみるなら重回帰分析\index{じゅうかいきぶんせき@重回帰分析}、経時的な変化を調べるなら反復測定分散分析\index{はんぷくそくていぶんさんぶんせき@反復測定分散分析}など、ある程度やるべきことは決まっている。これらのうち、分散分析、生存時間解析、重回帰分析、反復測定分散分析については、EZR を使ってメニュー操作で分析でき、保健学共通特講 IV, VIII のテキストで、ある程度説明してあるので、そちらを参照されたいが、非線形回帰は EZR でサポートされていないので、LD50 や ED50 の推定法については、本テキストの応用回帰分析の中で説明する。

なお、実験データについて統計解析をされる方に対して素晴らしいパースペクティブを与えてくれる本として、三中信宏 (2015)『みなか先生といっしょに 統計学の王国を歩いてみよう〜情報の海と推論の山を越える翼をアナタに！』羊土社を読むことをお勧めする。

The data from field survey

In the field survey, through the combination of quantitative measurements, questionnaire and interview, the data may include both continuous and categorical variables. フィールド調査\index{ふぃーるどちょうさ@フィールド調査}をすると、質問紙と測定の両方を実施することが珍しくない。縦断研究\index{じゅうだんけんきゅう@縦断研究}の場合には連結可能匿名化\index{れんけつかのうとくめいか@連結可能匿名化}が必要だが、同一人をどうやって追跡するかが大きな問題となる。あらゆるタイプのデータが含まれる可能性があり、データ解析もある程度探索的にならざるを得ないので、統計解析としては最も難しい。しかも、欠損値\index{けっそんち@欠損値}が珍しくないので、まずは欠損の質の検討が必要である。ランダムな欠損なら問題はないが、調べたい内容と欠損になるかどうかが関連していると非常にまずい。ランダムな欠損の場合、多重代入法\index{たじゅうだいにゅうほう@多重代入法}(multiple imputation)によって欠損値を補うことが良く行われる。\verb!mice!\index{mice}パッケージや\verb!Amelia!\index{Amelia}パッケージを使うことが多い\footnote{多重代入法については、高橋将宜・渡辺美智子(2017)『欠測データ処理：R による単一代入法と多重代入法』共立出版、ISBN: 978-4-320-11256-8 をお薦めする。なぜ欠損値を含むケースを単純にすべて除去したり、単一代入で済ませることが

バイアスにつながってしまうのか、実際に多重代入をする際にどこをチェックしなくてはいけないのか、など明確に解説した素晴らしいテキストである。}。

最も大切なのは、{\bf 他の解析をする前に、データの分布\index{でーたのぶんぷ@データの分布}をよく見ておく}ことである。カテゴリ変数なら度数分布図\index{どすうぶんぷず@度数分布図}（\verb!barplot()!\index{barplot()}で描くことができる）、量的変数ならヒストグラム\index{ひすとぐらむ@ヒストグラム}（\verb!hist()!\index{hist()}で描くことができる）や正規確率プロット\index{せいきかくりつぷろっと@正規確率プロット}（\verb!qqnorm()!\index{qqnorm()}で描くことができる）を作るのが常道である。健診データでは血圧正常値とかメタボリックシンドロームの腹囲カットオフ値のように、連続量として測定した値を正常・異常の 2 値情報にしてしまうことが良く行われるが、境界付近の値を単純に 2 値化\index{2 ちか@2 値化}することは問題がある。分布が明らかに二峰性\index{にほうせい@二峰性}なら谷のところで区分することに問題はないが、正規分布\index{せいきぶんぷ@正規分布}に近い形をしていて、質的な違いがあるわけでもないのに、固定されたカットオフ値を使って 2 値情報にしてしまうことは薦められない。統計解析のセンスからすれば、そのような場合は連続量のまま扱う方が筋が良い。どうしてもカテゴリ化したければ、明らかな低値、中間値、明らかな高値というカテゴリにして、明らかな低値と明らかな高値の 2 群間で比較することを検討すべきである。

2 変数の関連を分析する場合、どちらもカテゴリならモザイクプロット\index{もざいくぷろっと@モザイクプロット}（\verb!mosaicplot()!\index{mosaicplot()}）、片方がカテゴリでもう片方が量ならカテゴリ変数で層別したストリップチャート\index{すとりっぷちゃーと@ストリップチャート}（\verb!stripchart()!\index{stripchart()}で描ける）や箱ひげ図\index{はこひげず@箱ひげ図}（\verb!boxplot()!\index{boxplot()}で描ける）、どちらも量なら散布図\index{さんぷず@散布図}（\verb!plot()!\index{plot()}で描ける）を作る。

3 変数以上の場合は、3 つめ以降の変数は色やプロット記号を変えるなどして 2 次元グラフの重ね描きとして表現するか、3 つめ以降の変数で層別して複数の 2 次元グラフを作成するなど、さまざまな手法がある（詳しくは後述）。

Preprocessing the data using R

R で使うデータは、通常、表形式で入力したデータフレーム\index{でーたふれーむ@データフレーム}になる。原則として、1 個体が 1 行になるように作成する。異なる時点での測定値や、複数選択の選択肢は、別々の変数（列）にする。1 行目は変数名にする。変数名はアルファベットで始まるようにし、英数字とピリオドだけからなるようにすべきである。グラフの軸ラベルを漢字で表記したい場合は、グラフ描画関数の中で指定すべきであり、変数名は英数ピリオドだけにする方がエラーが起きにくい。

前処理が必要な場合が多々あるのでまとめておく。

Long format and wide format

10 人の被験者がいて、コーヒーを飲む前後で百マス計算をしてもらい、誤答数を記録した結果が、以下のように得られているとする。

\begin{tabular}{lccccccccc}
\hline
被験者 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10\cr
\hline
飲用前 & 5 & 3 & 2 & 7 & 3 & 1 & 4 & 3 & 9 & 3\cr
飲用後 & 4 & 3 & 1 & 6 & 2 & 2 & 2 & 2 & 5 & 2\cr
\hline
\end{tabular}

\begin{figure}[ht]
\begin{center}
\includegraphics[width=6cm]{coffeetest.pdf}
\end{center}
\caption{コーヒー飲用前後での百マス計算の誤答数の変化\label{fig:coffeedata}}
\end{figure}

もちろん Excel や LibreOffice Calc などで図\ref{fig:coffeedata}のように入力してから、範囲選択してコピーし、

\begin{screen}\begin{verbatim}
coffee <- read.delim("clipboard")
\end{verbatim}\end{screen}

のように\footnote{MacOS では\verb!read.delim(pipe("pbpaste"))!としなくてはいけない。Windows には\verb!"clipboard"!という名前のデバイスがあるが、MacOS では\verb!"pbpaste"!というアプリにパイプ処理でリダイレクトする。Windows にも\verb!"clip.exe"!というアプリがあるので、同様にパイプ処理にすることもできるはずである。誰もそんなことはしないが。}、データフレーム\verb!coffee!に付値しても

よいし、タブ区切りまたはコンマ区切りのテキストファイル、例えば\verb!e:/work/coffeedata.txt!として保存し、それを

\begin{screen}\small\begin{verbatim}
coffee <- read.delim("e:/work/coffeedata.txt")
\end{verbatim}\end{screen}

のようにして読み込んでもよい。

しかし、この程度のデータの量ならば、次の R コードとして直接ベクトルを定義し（\verb!data.frame()!の中では\verb!<-!でなく\verb!=!を使うことに注意。つまり、ここでやっているのはオブジェクトへの付値ではなく、ラベル付けである）、データフレームとして付値する方が簡単である。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/coffee.R(1)}\small\begin{verbatim}
coffee <- data.frame(
 pid = 1:10,
 pre = c(5, 3, 2, 7, 3, 1, 4, 3, 9, 3),
 post = c(4, 3, 1, 6, 2, 2, 2, 2, 5, 2))
\end{verbatim}\end{itembox}

コーヒー飲用前後で誤答数が統計学的に有意に変化したかどうか知りたい場合は、この形のまま、以下の枠内のコードを打てば同じ人の誤答数を線でつないだグラフが描かれ、検定もできる。コーヒー飲用後、百マス計算の誤答数が有意水準 5\%で統計学的に有意に減ったことがわかる。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/coffee.R(2)}\small\begin{verbatim}
plot(c(1, 2), c(0, 10), type="n", frame=FALSE, axes=FALSE,
 xlab="コーヒー飲用", ylab="誤答数")
segments(1, coffee$pre, 2, coffee$post)
axis(1, 1:2, c("前", "後"))
axis(2, 0:10, 0:10)
t.test(coffee$post, coffee$pre, paired=TRUE)
# 前後の差の母平均が 0 という検定と同じなので次の行でも同じ結果
# t.test(coffee$post-coffee$pre, mu=0)
\end{verbatim}\end{itembox}

ここで、仮に個人差を無視し、コーヒーを飲んでいない群と飲んだ群とで誤答数を比較するという操作をしたいときは、データの形を変える必要がある。簡単に言えば、次の枠内のようにしてデータを積み上げるとよい。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/coffee.R(3)}\small\begin{verbatim}
scoffee <- data.frame(
 pid = rep(coffee$pid, 2),
 errors = c(coffee$pre, coffee$post),
 setting = factor(c(rep(1, 10), rep(2, 10)), labels=c("pre", "post")))
\end{verbatim}\end{itembox}

積み上げ型データは、もっと簡単に、

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/coffee.R(4)}\small\begin{verbatim}
scoffee2 <- stack(list(pre=coffee$pre, post=coffee$post))
\end{verbatim}\end{itembox}

でも作成できる。ただし\verb!pid!は引き継がれないし、数値変数名は\verb!values!、グループ変数名は\verb!ind!と固定されている。同様に、\verb!car!パッケージの\verb!reshape()!関数を使えば、縦長形式と横長形式を相互変換できる。ただし、変数名としてピリオドの後に時点を示す数値を含んでいる必要がある。この場合、数値変数名は\verb!t!、グループ変数名（時点名）は\verb!time!と固定されている。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/coffee.R(5)}\small\begin{verbatim}
library(car)
colnames(coffee) <- c("pid", "t.0", "t.1") # pre→t.0、post→t.1 が必須
scoffee3 <- reshape(coffee, direction="long",
 idvar="pid", varying=c("t.0","t.1"))
# coffee3 <- reshape(scoffee3, direction="wide") # で戻せる
\end{verbatim}\end{itembox}

このようにして作った積み上げ型データ\verb!scoffee!を使って 2 群間の平均値の比較をするには以下のようにする。ストリップチャートが描かれ、Welch の方法による等分散性を仮定しない t 検定が実行される。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/coffee.R(6)}\small\begin{verbatim}
stripchart(errors ~ setting, data=scoffee, method="jitter",
 vert=TRUE, ylim=c(0, 10))
meanerrors <- tapply(scoffee$errors, scoffee$setting, mean)
sderrors <- tapply(scoffee$errors, scoffee$setting, sd)
igroups <- c(1.1, 2.1)
points(igroups, meanerrors, pch=18, cex=2)
arrows(igroups, meanerrors-sderrors, igroups, meanerrors+sderrors,
 angle=90, code=3)
t.test(errors ~ setting, data=scoffee)
\end{verbatim}\end{itembox}

ノンパラメトリックな図示と検定の場合はもっと簡単で、

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/coffee.R(7)}\small\begin{verbatim}
plot(errors ~ setting, data=scoffee)
wilcox.test(errors ~ setting, data=scoffee)
\end{verbatim}\end{itembox}

とするだけで層別箱ヒゲ図が描かれ（\verb!setting!という変数がファクター型なので自動的に\verb!boxplot()!が呼ばれる）、ウィルコクソンの順位和検定（マン＝ホイットニーの U 検定と数学的に同一）が実行される。

t検定でもウィルコクソンの順位和検定でも、このように個人差を無視してしまうと、このデータでは 2 群間には統計学的な有意差を見いだすことができなくなることがわかる。従って、あくまでデータの性質

に従ってデータファイルを設計すべきであり、このように積み上げ操作をすることは必ずしも一般的でないが、データフレームの前処理として覚えておくと役に立つことがある。

Table (data.frame) manipulation

カテゴリデータを表形式で操作するテクニックを、簡単な例で示す。\url{https://minato.sip21c.org/medstat/sample11.txt}は 40 人分の、年齢\verb!AGE!、曝露の有無\verb!EXPOSURE!（\verb!YES!と\verb!NO!の 2 値）、疾病の有無\verb!DISEASE!（\verb!YES!と\verb!NO!の 2 値）からなるタブ区切りテキストデータである。これを\verb!dat!というデータフレームに読み込むには、

\begin{screen}\begin{verbatim}
dat <- read.delim("https://minato.sip21c.org/medstat/sample11.txt")
\end{verbatim}\end{screen}

とする。このデータについてのさまざまな集計方法をまとめてみる。

\begin{description}
\item[EXPOSURE の集計] \verb!table(dat$EXPOSURE)!と打てば、以下が表示される。\par
\begin{screen}\begin{verbatim}
NO YES
20 20
\end{verbatim}\end{screen}
\item[結果を度数分布表ベクトルとしてオブジェクト EXC に付値] \verb!EXC <- table(dat$EXPOSURE)!
\item[DISEASE の集計] \verb!table(dat$DISEASE)!
\begin{screen}\begin{verbatim}
NO YES
16 24
\end{verbatim}\end{screen}
\item[曝露ありの人の DISEASE の集計] \verb!table(dat$DISEASE[dat$EXPOSURE=="YES"])!
\begin{screen}\begin{verbatim}
NO YES
 4 16
\end{verbatim}\end{screen}
\item[曝露あり結果を EXD に付値] \verb!EXD <- table(dat$DISEASE[dat$EXPOSURE=="YES"])!
\item[曝露なし結果を NED に付値] \verb!NED <- table(dat$DISEASE[dat$EXPOSURE=="NO"])!
\item[２つのオブジェクトを行方向に結合] \verb!rbind(NED, EXD)!で曝露の有無と疾病の有無のクロス集計結果が得られる。
\begin{screen}\begin{verbatim}
    NO YES
NED 12   8
EXD 4 16
\end{verbatim}\end{screen}
\item[クロス集計] 実はいきなり\verb!table(dat$EXPOSURE,dat$DISEASE)!でクロス集計できる。
\begin{screen}\begin{verbatim}
      NO YES

```
        NO  12   8
        YES  4  16
\end{verbatim}\end{screen}
\item[表題付きクロス集計] \verb!xtabs(~EXPOSURE+DISEASE, data=dat)!
\begin{screen}\begin{verbatim}
             DISEASE
       EXPOSURE NO YES
          NO  12   8
          YES  4  16
\end{verbatim}\end{screen}
\item[行列定義] 各組合せ人数が最初からわかっていれば、\verb!X <- matrix(c(12, 4, 8, 16), 2, 2)!
\item[ラベルをつける] \verb!rownames(X) <- c("NO", "YES"); colnames(X) <- c("NO", "YES")!
\item[ラベル(2)] \verb!dimnames(X) <- list(c("非曝露", "曝露"), c("健康", "病気"))!
\item[テーブルにする] \verb!attr(X, "class") <- "table"!
\item[独立性のカイ二乗検定] \verb!chisq.test(X)!
\item[Fisher の正確確率検定] \verb!fisher.test(X)!
\item[年齢 60 歳以上/未満の 2 群に区分した変数 AC を dat 内に作る] 次のどちらかを実行する。以下の説明では\verb!ifelse()!を使ったとする。
\begin{screen}\begin{verbatim}
dat$AC <- cut(dat$AGE, c(min(dat$AGE), 60, max(dat$AGE)+1),
 right=FALSE)
dat$AC <- factor(ifelse(dat$AGE<60, 1, 2),
 labels=c("<60", "60<="))
\end{verbatim}\end{screen}
\item[AC で元データを 2 群に分け、2 群別々にクロス集計\index{くろすしゅうけい@クロス集計}して YTAB と ETAB に付値] 以下のようにする。
\begin{screen}\begin{verbatim}
YTAB <- xtabs(~EXPOSURE+DISEASE, data=subset(dat,AC=="<60"))
ETAB <- xtabs(~EXPOSURE+DISEASE, data=subset(dat,AC=="60<="))
\end{verbatim}\end{screen}
\item[60 歳未満/以上で別々に Fisher の正確確率検定] \verb!fisher.test(YTAB); fisher.test(ETAB)!
\item[3 次元のクロス表を作る] \verb!D3TAB <- array(c(YTAB, ETAB), dim=c(2,2,2))!とすると、3 次元のクロス表が\verb!D3TAB!にできる（ラベルが全部消えてしまうが）。\verb!D3TAB!と打つと、次のように見える。
\begin{screen}\begin{verbatim}
      , , 1

          [,1] [,2]
      [1,]    4    3
      [2,]    4   13

      , , 2

          [,1] [,2]
      [1,]    8    5
      [2,]    0    3
\end{verbatim}\end{screen}
```

\item[xtabs や table で作る] 実は以下のどちらかで直接 3 次元クロス表ができる。
\begin{screen}\begin{verbatim}
D3TAB <- xtabs(~EXPOSURE+DISEASE+AC, data=dat)
D3TAB <- table(dat$EXPOSURE, dat$DISEASE, dat$AC)
\end{verbatim}\end{screen}
\item[3 次元の表から年齢層別に二次元クロス集計表を取りだす] 3 次元クロス表から 2 次元クロス表を取り出すには、
\begin{screen}\begin{verbatim}
YTAB <- D3TAB[,,1]
ETAB <- D3TAB[,,2]
\end{verbatim}\end{screen}
\item[60 歳未満/以上どちらでも曝露と疾病に関連はないという帰無仮説の検定] マンテルヘンツェルの検定\footnote{\verb!mantelhaen.test(D3TAB)!}を行うのが普通である。帰無仮説が有意水準 5\%で棄却されるので、どの年齢層でもこの曝露と疾病の間には統計学的に有意な関連があるといえる。また共通オッズ比は\verb!7.3 [1.29, 41.6]!であり、年齢で層別した場合に、どの年齢層でも共通して非曝露群に比べて曝露群での疾病オッズが 7.3 倍と見ることができる。
\item[3 次の交互作用がない帰無仮説の Woolf の検定] vcd ライブラリに入っていて以下で実行できる。
\begin{screen}\begin{verbatim}
library(vcd)
woolf_test(D3TAB)
\end{verbatim}\end{screen}
\end{description}

Recode and character manipulation

Recoding the categorical data

例えば、\verb!x!というデータフレームの\verb!AREA!という数値変数（値は 1〜9）に地域区分が入っている状態を考えよう。次のコードで生成できる。

\begin{screen}\begin{verbatim}
set.seed(54321) # 擬似乱数列に初期値 54321 を与える
x <- data.frame(AREA=sample(1:9, 100, replace=TRUE))
\end{verbatim}\end{screen}

AREA を地域名（A〜I）がついたファクター型に変換し、かつ 3 種類の街区（市街地=A,C,G、農村部=B,F,H、工業地区=D,E,I）に区分し直した新しい分類変数\verb!REG!を作って同じデータフレームに入れたいときは、次のようにする。

\begin{screen}\begin{verbatim}
NAREA <- c('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I')
# 実は LETTERS[1:9]と同じ
x$AREA <- factor(x$AREA, labels=NAREA)
x$REG <- factor(ifelse(x$AREA %in% c('A', 'C', 'G'), 1,
 ifelse(x$AREA %in% c('B', 'F', 'H'), 2, 3)),

```
 labels=c('市', '農', '工'))
# 以下は別解
NREG <- c('市', '農', '市', '工', '工', '農', '市', '農', '工')
x$REG <- NREG[as.integer(x$AREA)]
```

Manipulating character strings

R 本体が最も苦手とする処理の 1 つが文字列操作\index{もじれつそうさ@文字列操作}である。以下、いくつかの役に立つ操作をまとめておく。なお、本格的に文字列操作をしたい場合は、\verb!stringr!や\verb!stringi!といった文字列処理用のパッケージを用いると良いらしい\footnote{\url{https://rpubs.com/uri-sy/demo_stringr}や\url{https://qiita.com/kohske/items/85d49da04571e9055c44}を参照。}。

\begin{description}
\item[ファイルからの読み込みで文字列をファクターに自動変換させない] \verb!read.delim()!\index{read.delim()}関数などで、文字列をファイルから読み込むとき、通常は自動的にファクター型になる。この自動変換をさせないグローバルオプションが\verb!options(stringsAsFactor=FALSE)!\index{options(stringsAsFactor=FALSE)}である……というのが R-3.6.3 までの仕様だったが、R-4.0.0 以降、この自動変換はしないのがデフォルトになった。自動変換したいときは、\verb!read.delim()!、\verb!read.csv()!などの関数の中に、オプションとして\verb!stringsAsFactors=TRUE!を入れる必要がある。
\item[データフレーム内のファクターを文字列に] データフレーム\verb!bob!の中のファクター型の変数を一括で文字列型\index{もじれつがた@文字列型}に変えたい場合は以下のようにする。
```
i <- sapply(bob, is.factor)
bob[i] <- lapply(bob[i], as.character)
```
\item[数値を書式付きで文字列に変換] C 言語と同様の仕様で\verb!sprintf()!\index{sprintf()}という関数が使える。表示桁長を見やすく揃えるときも便利。例えば、\verb!sprintf("%09d", 4)!の結果は以下。図中などに浮動小数点表示をさせたくないときにも便利。
```
> sprintf("%09d", 4)
[1] "000000004"
> x <- 123456789012345
> x
[1] 1.234568e+14
> sprintf("%15.0f", x)
[1] "123456789012345"
> x <- 0.0000000456
> x
[1] 4.56e-08
> sprintf("%10.8f", x)
[1] "0.00000005"
```

\item[文字列処理関数群] \verb!paste()!\index{paste()}、\verb!substr()!\index{substr()}、\verb!strsplit()!\index{strsplit()}などであるが、あまり機能は充実していない。\verb!stringr!\index{stringr}パッケージを使うと、例えば、ある文字列に含まれる別の文字列の個数を返す\verb!str_count()!\index{str_count()---stringr}関数などが使える\footnote{例えば、\verb!str_count("abc1234def5432", "4")!は、第二引数の文字列が第一引数に 2 回出現するので 2 を返す。}。最近では\verb!stringi!パッケージを薦める人も多い\footnote{\url{https://qiita.com/kohske/items/85d49da04571e9055c44}など。}。
\end{description}

Various graphs drawn by R

{\bf 作図はデータ解析の常道}である。どんなに複雑な統計解析をする場合にも、データの性状を知るために作図は必須である。R では多種多様なデバイス（ベクトルグラフィックス＝図形ファイルとしてウィンドウズメタファイルや pdf、ポストスクリプトなど、ラスターグラフィックス＝画像ファイルとして tiff や jpeg など、あるいはコンピュータのディスプレイ）に作図することが可能だし、図形ファイルは後で PowerPoint、LibreOffice Draw/Impress に読み込んで「切り離す」ことで線単位で再編集でき、画像ファイルは Photoshop などのフォトレタッチソフトで加工できる。なお、ラスターグラフィックスデバイスの中では、bg="transparent"として背景に透過色を指定できる\verb!png()!も使いやすいデバイスである。データが何千点もある散布図など、ベクターグラフィックスよりもラスターグラフィックスにした方が操作が軽くなるしファイルサイズも小さくなる。

The basic process of drawing graphs

R の作図の基本プロセスは以下のステップを踏む。なお、R のグラフィックスには\verb!base!の他に\verb!grid!というシステムがあり、\verb!grid!を使って探索的作図ができることでよく知られている\verb!ggplot2!というパッケージもよく使われているが、このテキストでは\verb!grid!は扱わない。\verb!ggplot2!について知りたい方は、開発者である Hadley Wickham 自身が書いた本を徳島大学の石田基広さんが翻訳した、H.ウィッカム（著）、石田基広、石田和枝（訳）『グラフィックスのための R プログラミング：ggplot2 入門』シュプリンガー・ジャパン株式会社、ISBN 978-4-431-10250-2 を参照されたい。

\begin{enumerate}
\item \verb!pdf("ファイル名", width=横幅, height=高さ)!、\verb!win.metafile("ファイル名", width=横幅, height=高さ)!、\verb!windows()!のようにしてグラフィックスデバイスを開く。\verb!windows()!デバイスでもサイズ指定は可能である。省略するとインタラクティブに操作しているときはコンピュータのディスプレイ（OS が Microsoft Windows なら\verb!windows()!デバイスを開くのと同じ）、バッチ処理では pdf デバイスとして\verb!Rplot.pdf!というファイルが出力先になる（既に\verb!Rplot.pdf!が存在する場合は、上書きではなく、\verb!Rplot01.pdf!などと自動的に数字が加わったファイルができていくはずである）。
\item \verb!layout()!、\verb!par()!などで、そのデバイス上へのグラフの配置や余白を設定する。例えば\verb!layout(1:2)!とするとデバイスが上下 2 分割される\footnote{\verb!layout(matrix(c(1, 1, 2, 3), 2, 2)!とすると、デバイス左半分が第 1 のグラフ、右上が第 2 のグラフ、右下が第 3 のグラフを描く領域として分割される。つまり、m 行 n 列の行列を\verb!layout()!に渡し、中身の数字を描画順序を示す整数にすれば良いので、かなり複雑な画面分割もできる。もちろん画面分割などせず、別々に作ったグラフを、後で LibreOffice Draw などで加工しても良いわけだが、コードで書いておけば手作業がないので再現や修正が容易である。}。\verb!par()!でよく使われるのは、\verb!cex=2!によって文字とシンボルのプロットサイズを標準の 2 倍にするとか、\verb!family="sans"!でフォントをサンセリフ体にするとか\footnote{日本語を描画に使うときもこの\verb!family=!オプションは重要。}、\verb!las=1!で軸目盛ラベルが常に水平に書かれるようにするとか\footnote{これを指定しないと、縦軸の目盛ラベルは自動的に左 90 度回転される}、\verb!mar=c(4, 3, 3, 1)+0.1!として余白を 1 列ずつデフォルト値より狭くする（指定順序は下、左、上、右）といったオプションである。

\item \verb!plot()!や\verb!hist()!などの座標系設定を伴うメイングラフ描画関数でグラフを描く。\verb!xlim=c(横軸最小値, 横軸最大値)!で座標系の横軸、\verb!ylim=c(縦軸最小値, 縦軸最大値)!で座標系の縦軸を指定できる。\verb!log="x"!オプションをつけると横軸のみ対数軸になり、\verb!log="xy"!とすると両対数グラフになる。\verb!xlab="横軸のラベル"!、\verb!ylab="縦軸のラベル"!というオプションで軸ラベルを付けることができる。なお、\verb!plot()!で外枠を描きたくない場合は\verb!frame=FALSE!オプション、軸をカスタマイズしたい場合は\verb!axes=FALSE!オプションを付ける。座標系は設定したいけれどもデータをプロットしたくない場合は、\verb!type="n"!オプションを付ける。
\item \verb!axes=FALSE!だった場合は、\verb!axis(1, 数値ベクトル, ラベル文字列ベクトル)!で横軸、\verb!axis(2, 数値ベクトル, ラベル文字列ベクトル)!で縦軸を設定する（3 で上、4 で右にも軸を付けられる）
\item \verb!lines()!や\verb!arrows()!や\verb!text()!や\verb!legend()!でグラフに追記する
\item \verb!dev.off()!でデバイスが閉じられ、描画が完了する
\end{enumerate}

The functions to draw the main graphs

\begin{description}
\item[hist()] ヒストグラムを描く
\item[qqnorm()] 正規確率プロットを描く
\item[barplot()] 棒グラフを描く。行列（＝2 次元クロス集計表）を与えると、積み上げ棒グラフやサブグループ別の棒グラフ（\verb!beside=TRUE!オプションを付けた場合。デフォルトは\verb!FALSE!なので積み上げ棒グラフになる）が描ける。\verb!horiz=TRUE!にすると横棒グラフになる（デフォルトは\verb!horiz=FALSE!なので縦棒グラフになる）
\item[boxplot()] 箱ひげ図を描く
\item[stripchart()] ストリップチャートを描く
\item[dotchart()] ドットチャートを描く
\item[mosaicplot()] モザイクプロットを描く
\item[pie()] 円グラフを描く
\item[plot()] \verb!plot()!は総称的な関数なので、与えるオブジェクトによって動作が変わる。2 つのカテゴリ変数をコンマで区切って与えればモザイクプロットになるし、\verb!plot(量的変数 ~ カテゴリ変数, data=データフレーム)!のようにするとカテゴリ変数で層別した層別箱ひげ図が描かれるし、2 つの量的変数をカンマで区切って与えるか、\verb!plot(量的変数 ~ 量的変数, data=データフレーム)!とすれば散布図が描かれる。\verb!x!というデータフレームに 2 つの量的変数\verb!A!と\verb!B!があるとき、\verb!plot(xA, xB)!でも\verb!plot(B ~ A, data=x)!のどちらでも、変数 A が横軸、変数 B が縦軸の散布図が描かれる。\verb!type="b"!とするとデータ点が線でつながれる。\verb!pch=!オプションでプロット記号を指定でき、\verb!col=!オプションで色を指定できる。
\item[pairs()] 複数の変数の同時散布図を描く
\item[matplot()] 複数の系列を 1 枚の散布図の中に重ね描きする
\item[coplot()] 第 3（＋第 4）の変数で層別した複数の散布図を描く。詳細は\verb!example(coplot)!で確認できるが、２つの要因で層別した同時散布図を\verb!coplot(y~x | a*b)!によって実行する場合、\verb!a!や\verb!b!が数値だと層別数は\verb!a!についても\verb!b!についてもデフォルトでは 6 である（\verb!numbers=!で変更可）。\verb!a!や\verb!b!がファクター型なら、カテゴリごとに\verb!plot(y~x)!される。
\item[dataEllipse()] \verb!car!パッケージが必要。散布図と集中楕円（確率楕円）を重ね描きする

\item[radarchart()] \verb!fmsb!パッケージが必要。レーダーチャート（蜘蛛の巣グラフ）を描く
\end{description}

Examples

以下、いくつかの事例について、具体的にグラフを作ってみる。

Drawing scattergram with different symbols for groups

散布図や層別ストリップチャートで第３の変数によってプロット記号を変えてみると、多くの情報が得られる。例えば、身長と体重の関係を散布図にするとき、男女別にプロット記号の形や色を変えると、男女込みにしたときに見られる相関関係は、男性が女性よりも身長も体重も平均して大きい傾向があることによって実際以上に強い正の相関関係があるように見えていることがわかる。

ここでは、X、Y、Z という３つの村があって、それぞれ身長と体重のデータがあって、その関係を村ごとにマークを変えてプロットしたいとする。データは、\url{https://minato.sip21c.org/advanced-statistics/v3hw.txt}からタブ区切りテキストとして入手できる。変数名は村が\verb!VG!、身長が\verb!HEIGHT!、体重が\verb!WEIGHT!である。データを\verb!x!というデータフレームに読み込み、まずざっくりと村ごとに分けた散布図を描くには\verb!coplot()!を使う。

\begin{itembox}[l]{https://minato.sip1c.org/advanced-statistics/scdehot.R(1)}\begin{verbatim}
URL <- "https://minato.sip21c.org/advanced-statistics/v3hw.txt"
v3 <- read.delim(URL, stringsAsFactors=TRUE)
plot(WEIGHT ~ HEIGHT, data=v3)
coplot(WEIGHT ~ HEIGHT | VG, data=v3)
\end{verbatim}\end{itembox}

$$\includegraphics[width=8cm]{coplotxyz.pdf}$$

これだと村落間の違いが分かりにくいので、村の名前をそれぞれ違う色で身長と体重の座標位置にプロットしてみる。コードは次の通り。

\begin{itembox}[l]{https://minato.sip1c.org/advanced-statistics/scdehot.R(2)}\begin{verbatim}
plot(WEIGHT ~ HEIGHT, data=v3, pch=as.character(VG),
 col=as.integer(VG), xlab="身長(cm)", ylab="体重(kg)",
 main="3 村落住民の身長と体重の関係")
\end{verbatim}\end{itembox}

$$\includegraphics[width=10cm]{v3hwplot-1.pdf}$$

村の名前をプロットするのは見栄えが悪いので、適当なシンボルを使ってプロットし、凡例を付記する方が良い。次のようにする。

\begin{itembox}[l]{https://minato.sip1c.org/advanced-statistics/scdehot.R(3)}\begin{verbatim}

plot(WEIGHT ~ HEIGHT, data=v3, pch=as.integer(VG),
 col=as.integer(VG), xlab="身長(cm)", ylab="体重(kg)",
 main="3 村落住民の身長と体重の関係")
series <- 1:length(levels(v3$VG))
legend("topleft", pch=series, col=series, legend=levels(v3$VG))
\end{verbatim}\end{itembox}

$$\includegraphics[width=10cm]{v3hwplot-2.pdf}$$

このように色とシンボルを組み合わせると多くの水準を描き分けることができる。\verb!pch!に与える値として、1 から 25 まではプロットとして適切なシンボルが既に定義されている（26 から 32 は空白で、33 以上は文字や記号）ので、\verb!col="red"!とか\verb!col="blue"!などと色を指定するか、剰余を使うなどして周期的変数を生成して色を変えれば 120 くらいは何とかなる。他にも、以下 2 つの方法がある。

\begin{itemize}
\item \verb!text()!関数を使って文字列を重ね打ちする: \verb!plot(x, y)!の後に（\verb!pch='.'!や\verb!pch=20!でプロット記号を小さい点にすると良い）、
\begin{screen}\small\begin{verbatim}
text(x, y, paste(string), pos=4, offset=0.5)
\end{verbatim}\end{screen}とすれば、\verb!string!を\verb!(x, y)!の点の右側に表示してくれる。
\item \verb!identify()!関数を使う: すべてのデータ点を特定する必要はないので、必要な点についてだけ情報を表示できるのがベストであろう。\verb!plot(x, y)!の後に\verb!identify(x, y, labels=string)!としておくと、プロットの後に十字型のマウスカーソルが出現するので、画面上で\verb!string!を表示したい点の上でクリックすれば\verb!string!が出現する。描画ウィンドウのメニューの\verb!stop!からか、右クリックメニューから\verb!stop!を選ぶまで複数の点をクリックできる。
\end{itemize}

ここまでやったなら、村落間で身長と体重の関係に違いがあるかどうかを知りたくなるだろう。集中楕円を描き、Hotelling の$\hbox{T}^2$検定を実行するには以下のコードを打つ。パッケージとして\verb!car!と\verb!Hotelling!が必要になるため、予めインストールしておく（たぶん既に入っていると思うので\verb!install.packages("car", dep=TRUE)!は必要ないのが普通であろうが、\verb!install.packages("Hotelling", dep=TRUE)!は必要な方が多いかもしれない）。Hotelling の$\hbox{T}^2$検定は、2 変量分布が 2 群間で異なるかどうかを調べるので、この場合のように 3 群あったら、2 群ずつ調べて、Holm の方法、FDR 法等で検定の多重性を調整せねばならない。以下のコードを実行すると、図\ref{fig:v3hwdataEllipse}が得られ、検定結果を見ると、Y 村と Z 村の間のみ有意水準 5\%で身長と体重の 2 変量分布に統計的に有意な差がある（\verb!p=0.011!）とわかる。

\begin{itembox}[l]{https://minato.sip1c.org/advanced-statistics/scdehot.R(4)}\small\begin{verbatim}
library(car)
dataEllipse(v3$HEIGHT, v3$WEIGHT, v3$VG, levels=0.8) #集中楕円描画
library(Hotelling)
Z <- split(v3[,c("HEIGHT","WEIGHT")], v3[,"VG"])
res12 <- hotelling.test(Z[[1]], Z[[2]])
res23 <- hotelling.test(Z[[2]], Z[[3]])
res31 <- hotelling.test(Z[[3]], Z[[1]])
res <- c(res12$pval, res23$pval, res31$pval)

```
names(res) <- c("X-Y", "Y-Z", "Z-X")
sort(res)*3:1 # Holm の方法で検定の多重性を補正
\end{verbatim}\end{itembox}

\begin{figure}[ht]
\begin{center}
\includegraphics[width=8cm]{v3hw-dataEllipse.pdf}
\end{center}
\caption{身長と体重の関係について 3 村落の 80\%確率楕円\label{fig:v3hwdataEllipse}}
\end{figure}
```

Visualization of the lifetable data by prefecture

厚生労働省のサイトで 2013 年 2 月 28 日に公開された、平成 22 年都道府県別生命表の概況(\url{https://www.mhlw.go.jp/toukei/saikin/hw/life/tdfk10/index.html})の「図表データのダウンロード」から Excel ファイル(\url{https://www.mhlw.go.jp/toukei/saikin/hw/life/tdfk10/dl/zuhyou.xls})をダウンロードして加工したデータを使って、都道府県別平均寿命の推移を示す折れ線グラフと、死因別損失余命の都道府県別プロファイルを示すレーダーチャートを、男女別に作成してみる。

男女別に都道府県別平均寿命の推移を示す（長野と沖縄だけ色を変えて強調した）折れ線グラフを描くコードは以下の通り。

```
\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/e0changes.R}\small\begin{verbatim}
e0 <- read.delim("https://minato.sip21c.org/demography/pref-e0-changes.txt",
 fileEncoding="CP932")
males <- t(e0[, 2:11])
colnames(males) <- e0$PREF
females <- t(e0[, 12:21])
colnames(females) <- e0$PREF
COL <- ifelse(e0$PREF=="長野", "blue",
 ifelse(e0$PREF=="沖縄", "pink", "lightgrey"))
LWD <- ifelse(e0$PREF=="長野", 2, ifelse(e0$PREF=="沖縄", 2, 1))
LTY <- ifelse(e0$PREF=="長野", 1, ifelse(e0$PREF=="沖縄", 1, 3))
years <- 1965+0:9*5
windowsFonts(JP1=windowsFont("MS Gothic"),
             JP2=windowsFont("MS Mincho"),
             JP3=windowsFont("Meiryo"),
             JP4=windowsFont("Biz Gothic"))

windows(width=1200, height=800) # for MacOS, quartz() can be used.
par(family="JP3") # to make pdf, family="Japan1" should be used.
# for MacOS, par(family="Japan1") should be used.
layout(t(1:2))
matplot(years, males, type="l", col=COL, lwd=LWD, lty=LTY,
 main="男性の都道府県別平均寿命の推移\n（青：長野、桃：沖縄、灰色：他）")
matplot(years, females, type="l", col=COL, lwd=LWD, lty=LTY,
 main="女性の都道府県別平均寿命の推移\n（青：長野、桃：沖縄、灰色：他）")
```

\end{verbatim}\end{itembox}

$$\includegraphics[width=12cm]{e0bypref.pdf}$$

このグラフから読み取れることはそれほど多くないが、1985 年までトップレベルだった沖縄男性の平均寿命が、1990 年から急に伸びが鈍化したこと、長野県男性も 1990 年までの伸びに比べると 1995 年以降は伸びが鈍化していることがわかる。女性については、男性と違って、最近まで沖縄の平均寿命の高さは他都道府県とは段違いだったのに、2005 年に追いつかれ、2005 年から 2010 年には横這いになってしまったことが一目で分かる。数値だけ眺めるよりわかりやすいと思う。

ちなみに、これは折れ線グラフなので、縦軸がゼロから始まっていないことに注意されたい。2010 年の男性の水準には、女性は 1980 年頃には既に到達していた。

\verb!NipponMap!パッケージの\verb!JapanPrefMap()!関数を使うと、都道府県別データからコロプレス図を作ることが簡単にできる。平成 22 年の都道府県別平均寿命を、ヒストグラムの階級を区分する Sturges アルゴリズムまたは\verb!pretty()!関数を使って適当に区分し、男女別に地図上で塗り分けるコードは以下である。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/e0Japan2010.R}
\scriptsize\begin{verbatim}
e0 <- read.delim("https://minato.sip21c.org/demography/pref-e0-changes.txt",
 fileEncoding="CP932")
mec <- cut(e0$e0M.2010, hist(e0$e0M.2010, plot=FALSE)$breaks, right=FALSE)
mec2 <- cut(e0$e0M.2010, pretty(e0$e0M.2010), right=FALSE)
fec <- cut(e0$e0F.2010, hist(e0$e0F.2010, plot=FALSE)$breaks, right=FALSE)
fec2 <- cut(e0$e0F.2010, pretty(e0$e0F.2010), right=FALSE)
mcol <- heat.colors(length(levels(mec)))[as.integer(mec)]
mcol2 <- heat.colors(length(levels(mec2)))[as.integer(mec2)]
fcol <- heat.colors(length(levels(fec)))[as.integer(fec)]
fcol2 <- heat.colors(length(levels(fec2)))[as.integer(fec2)]

windowsFonts(JP1=windowsFont("MS Gothic"),
        JP2=windowsFont("MS Mincho"),
        JP3=windowsFont("Meiryo"),
        JP4=windowsFont("Biz Gothic"))

windows(width=1200, height=800) # for MacOS, quartz() should be used.
par(family="JP4") # for MacOS, par(family="Japan1") should be used.
layout(matrix(1:4, 2, 2))

library(NipponMap)
JapanPrefMap(mcol, main="Life expectancy at birth in Japanese males in 2010")
legend("bottomright", fill=heat.colors(length(levels(mec))), legend=names(table(mec)))
JapanPrefMap(mcol2, main="2010 年日本人男性の都道府県別平均寿命\n(pretty による区切り)")
legend("bottomright", fill=heat.colors(length(levels(mec2))), legend=names(table(mec2)))
JapanPrefMap(fcol, main="Life expectancy at birth in Japanese females in 2010")
legend("bottomright", fill=heat.colors(length(levels(fec))), legend=names(table(fec)))

```
JapanPrefMap(fcol2, main="2010 年日本人女性の都道府県別平均寿命\n(pretty による区切り)")
legend("bottomright", fill=heat.colors(length(levels(fec2))), legend=names(table(fec2)))
\end{verbatim}\end{itembox}

$$\includegraphics[width=12cm]{e0Japan2010.pdf}$$
```

いくつかの指標をプロファイルとして多角形で示すのがレーダーチャートである。R では\verb!fmsb!パッケージに\verb!radarchart()!関数として実装してある。このデータから都道府県別死因別損失余命プロファイルを（やはり長野県と沖縄県を強調して）描くコードは下記の通り。

```
\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/cdradar.R}\small\begin{verbatim}
x <- read.delim("https://minato.sip21c.org/demography/pref-LLY-h22.txt",
 fileEncoding="CP932")
COL <- ifelse(x$PREF=="長野", "blue",
 ifelse(x$PREF=="沖縄", "pink", "lightgrey"))
LWD <- ifelse(x$PREF=="長野", 2, ifelse(x$PREF=="沖縄", 2, 1))
LTY <- ifelse(x$PREF=="長野", 1, ifelse(x$PREF=="沖縄", 1, 3))
VX <- c("悪性新生物","高血圧を除く\n 心疾患","脳血管疾患","三大死因",
 "肺炎","不慮の事故","交通事故\n（再掲）","自殺","腎不全","肝疾患",
 "糖尿病","高血圧","結核")
males <- x[,2:14]
females <- x[,15:27]
require(fmsb)
windowsFonts(JP1=windowsFont("MS Gothic"),
             JP2=windowsFont("MS Mincho"),
             JP3=windowsFont("Meiryo"),
             JP4=windowsFont("Biz Gothic"))

windows(width=1200, height=800) # for MacOS, quartz() should be used.
par(family="JP4") # for MacOS, par(family="Japan1") may be used.
layout(t(1:2))
radarchart(males, maxmin=FALSE, pcol=COL, axistype=2, pty=32, plty=LTY,
 plwd=LWD, vlabels=VX,
 title="男性の死因別損失余命（平成 22 年度）\n（青：長野、桃：沖縄、灰：他都道府県）")
radarchart(females, maxmin=FALSE, pcol=COL, axistype=2, pty=32, plty=LTY,
 plwd=LWD, vlabels=VX,
 title="女性の死因別損失余命（平成 22 年度）\n（青：長野、桃：沖縄、灰：他都道府県）")
\end{verbatim}\end{itembox}

$$\includegraphics[width=12cm]{cdsllybypref.pdf}$$
```

このグラフはいろいろなことを示唆してくれる。一見してわかることは、平均寿命が男女とも最長の長野県は、男女とも、がんと肺炎による死亡が少ないということだ。一方、脳血管疾患によって失われている余命は比較的大きい。これは、長野県の人は漬け物をよく食べるため、元々塩分摂取量が多く、そのために脳卒中が多かったのを、食生活改善推進員さんが歩き回って塩分摂取量を減らし、そのおかげで脳卒中が減ったと言われているのだが、それでもまだ塩分摂取が高いということかもしれない。ただし、くも膜下出血のリスク因子としては遺伝も大きいので、塩分摂取だけが問題とは言い切れないが。なお、長野県

では、男性のみ交通事故によって失われている余命が大きいが、これは子供の交通事故死だと思われる。細くて見通しが悪くて歩道が狭い道路が多いのに外遊びする子供は多いので、飛び出しによる交通事故が比較的多いのであろうことは想像に難くない。沖縄のプロファイルから目立つのは、肝疾患、糖尿病が高いことだ。たぶん飲酒が多いせいだろう。女性のみ結核による損失余命が大きかったが、これは流行があったのかもしれない。

Comparison of 2 time-series data

世間では、時系列の２つの変数の推移グラフを重ねて、動きが似ているから関係があるとするロジックが使われることがある。例えば、生活クラブ大阪の 2016 年 2 月の「クラブ通信 Vol.93」の記事（\url{https://osaka.seikatsuclub.coop/excludes/osaka/img/member/bulletin/2016club/club.93.pdf}）では、年次を横軸、日本人一人あたりの年間牛肉消費量と子宮体がん発生数を縦軸にとって、前者を棒グラフ、後者を折れ線グラフとして重ね描きして、推移が似ているから関連があるのだと論じている。

日本人一人当たりの年間牛肉消費量は食糧需給表（\url{https://www.e-stat.go.jp/SG1/estat/List.do?lid=000001131797}）から、3-7 の中の牛肉というところから Excel のワークシートをダウンロードでき、子宮体がん発生数は、がんセンター（\url{https://ganjoho.jp/professional/statistics/statistics.html}）の「２．罹患データ（全国推計値）」から Excel のワークシートをダウンロードできるので、それぞれ該当データを抽出してタブ区切りテキスト形式にしたものを\url{https://minato.sip21c.org/advanced-staitstics/beef-and-corpus-uteri-carcinoma.txt}に掲載した。数値からみると、当該グラフで使われている「日本人一人当たりの年間牛肉消費量」は国内消費仕向量の粗食料の値であり（リンク先データでは BEEFCC とした）、歩留まりが考慮されていない。むしろ 1 人当たり供給量（リンク先データでは BEEFSP とした）の方が摂取量には近いと考えられる。リンク先データでは子宮体がん発生数は CUCI とし、年次は YEAR とした。

このデータを読み込んで、生活クラブのサイトと同じものを再現するコードは下記の通りである。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/beefutecan.R(1)}
\small\begin{verbatim}
URL <- "https://minato.sip21c.org/beef-and-corpus-uteri-carcinoma.txt"
x <- read.delim(URL)
par(las=1, mar=c(3, 4, 4, 4))
y <- barplot(x$BEEFCC, col="red", ylim=c(0, 14),
 main="Crude beef consumption and corpus uteri carcinoma incidence")
lines(y, x$CUCI/600, col="skyblue", lwd=2)
axis(4, 0:8*5/3, labels=0:8*1000)
axis(1, y, labels=x$YEAR)
\end{verbatim}\end{itembox}

$$\includegraphics[width=12cm]{overlaygraph.pdf}$$

しかし、これが真の相関関係（ある程度の規則性をもって大小をともにする関係）であるならば、２つの変数間の散布図を描いて年次推移を矢印でつないだ場合に、矢印の傾きと全体の傾向が一致するはずである。こういう推移グラフを描くのも R ならば簡単である。ポイントは$[-1]$によってベクトルの最初の要

素を削除したベクトルを作るところで、それにさえ気づけば、\verb!arrows(x0, y0, x1, y1)!関数で\verb!(x0, y0)!から\verb!(x1, y1)!への矢印を追記できるので、推移グラフが完成する。

\begin{screen}\small\begin{verbatim}
データは同じなので読み込み部分は省略
URL <- "https://minato.sip21c.org/beef-and-corpus-uteri-carcinoma.txt"
x <- read.delim(URL)
plot(x$BEEFCC, x$CUCI, type="p", pch=16, xlab="年間牛肉消費量",
 ylab="年間子宮体がん発生数",
 main="日本人における牛肉消費量と子宮体がん発生数の年次推移")
arrows(x$BEEFCC, x$CUCI, c(x$BEEFCC[-1], NA), c(x$CUCI[-1], NA),
 col="navy", length=0.1)
\end{verbatim}\end{screen}

$$\includegraphics[width=10cm]{transitiongraph.pdf}$$

真の正の相関関係であれば、左下と右上を結ぶ方向に推移するはずだが、ほとんどそういう推移になっている年度はない。摂取から発症までの潜伏期間を考えてプロットする年をずらしても、きれいな関係にはならなそうなので、おそらく、これら２つの変数の間の相関は擬似相関と考えられる。

なお、このグラフがアクティブなグラフとして表示されている状態で、\verb!identify()!\index{identify}関数を以下のように実行すると、散布図上で点を選んで年の情報を表示させることができる。

\begin{screen}\small\begin{verbatim}
identify(x$BEEFCC, x$CUCI, x$YEAR, col="red")
\end{verbatim}\end{screen}

マウスカーソルが十字型になり、グラフ上の任意の描画点の近傍でクリックすれば、その年を赤い字で（\verb!col="red"!としているため）書き加えることができる（停止は右クリックから選ぶか、ウィンドウ左上の「停止」から可能）。

$$\includegraphics[width=10cm]{identify-example.pdf}$$

\verb!x$YEAR!のところを\verb!x$CUCI!とすれば、その点の子宮体がん発生数を数値として表示させることもできる。選択せずに、すべての点に年をオレンジ色で表示させたければ、以下のコードでできる（ごちゃごちゃするのでお勧めしないが）。

\begin{screen}\small\begin{verbatim}
text(x$BEEFCC, x$CUCI, x$YEAR, col="orange", pos=1)
\end{verbatim}\end{screen}

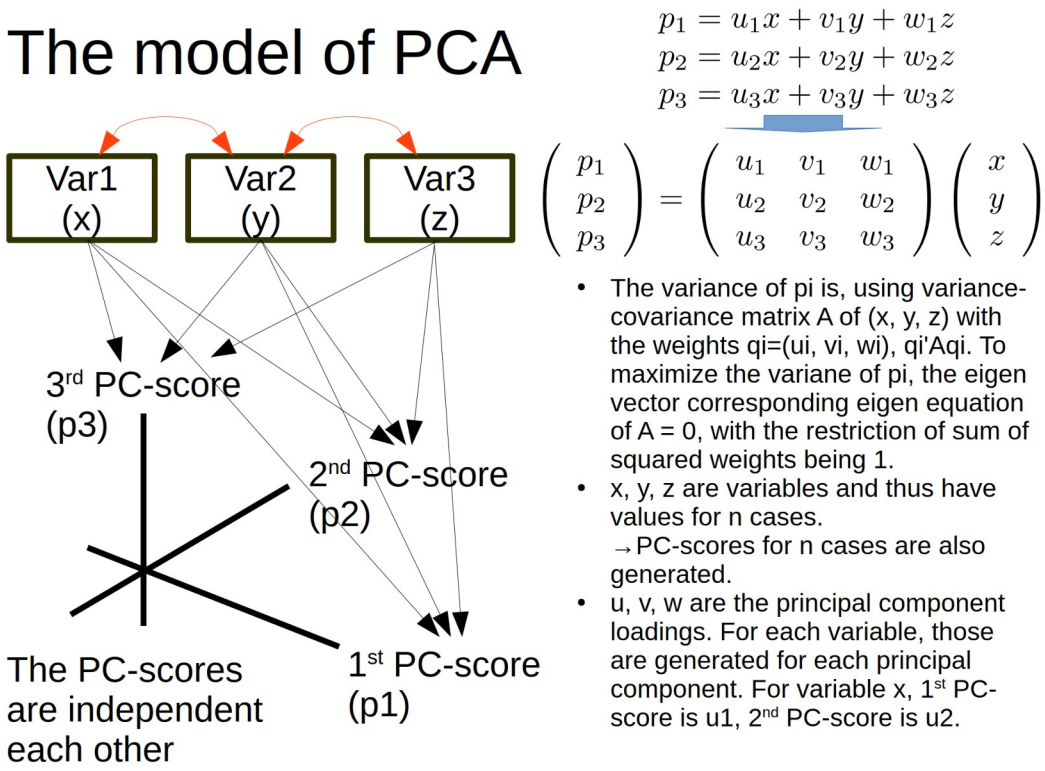
Factor Analysis

Factor analysis and principal component analysis

因子分析\index{いんしぶんせき@因子分析}とは、見かけは主成分分析\index{しゅせいぶんぶんせき@主成分分析}に似ているので混同されやすいが、指向性は真逆な分析法である。まずこれら２つを区別しよう。

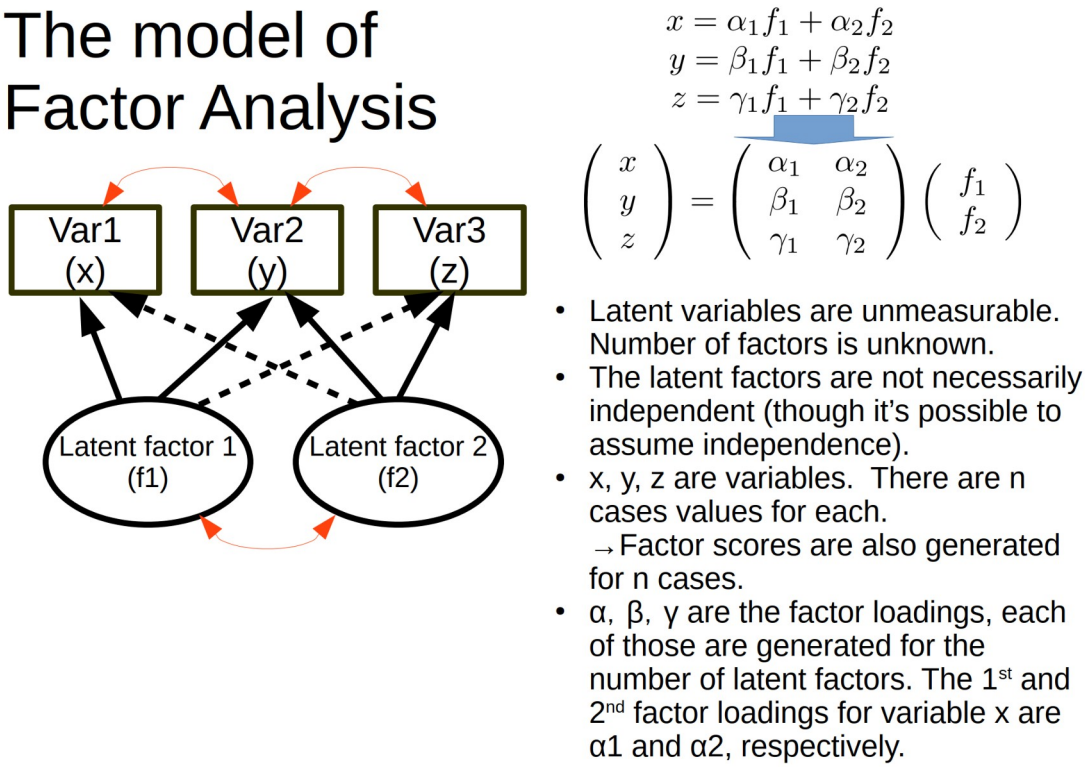
What is principal component analysis?

主成分分析においては、観測された多くの変数の分散を、{\bf それらの変数の線形結合として表される互いに独立な}主成分\index{しゅせいぶん@主成分}の合成ベクトルとして記述する。主成分は、元のデータがもつ全分散のうち、より多くの割合を説明する順に選択される。2 番目の主成分は、1 番目の主成分と独立という制約の下で、次に多くの割合を説明するように選ばれる。理想的な結果としては、少数の主成分によって元データの分散の大部分が説明され\footnote{Oxford Handbook for Medical Statistics, 4th Ed.には、通常、２つか３つの主成分で分散の少なくとも 80\%が説明される（即ち、第 3 主成分までで累積寄与率が 0.8 を超えるのが普通）、と書かれている。}、多くの変数によって高次元空間に位置づけられていた個々のデータ（人を対象として得られた測定値の場合は個人を示す）を、これら少数の主成分の得点によって張られる低次元空間で位置づけるという、{\bf 次元の縮小}を行うことができる（図\ref{fig:princomp}）。



What is factor analysis?

因子分析\index{いんしぶんせき@因子分析}は、図\ref{fig:factor}に示す通り、観測された変数（互いに関連をもっている）の背後にあるけれども観測不可能な潜在因子\index{せんざいいんし@潜在因子}を想定し、それら潜在因子の線形結合によって観測された変数を記述するモデルである。次のようにまとめられる。



\begin{figure}[ht]
\begin{center}
\includegraphics[width=10cm]{factor.pdf}
\end{center}
\caption{因子分析のモデル\label{fig:factor}}
\end{figure}

\begin{description}
\item[真面目な説明] 観察された変数の背後に隠れている因子を見いだすこと。この隠れた因子は直接測定できないが、観察された変数の「自然のグルーピング」になっている\footnote{データセット内のお互いに強く相関する変数のサブセットで、他の変数とは弱い相関をもつ。見つかった因子は、理論的に解釈可能な、隠れた「次元」に対応するはずである。}。
\item[実用的な説明] 互いに相関のある変数について、情報を集約して数を減らすこと。この意味では、主成分分析と似ている（向きは逆だが）。
\end{description}

Basic usage of PCA

R で主成分分析を行う関数には、\verb!princomp()!と\verb!prcomp()!がある。どちらも標準で含まれているので、追加パッケージは必要ない。ただし、群馬大学青木繁伸教授が\url{http://aoki2.si.gunma-u.ac.jp/R/src/pca.R}で公開している\verb!pca()!という関数の方が高機能であり、そちらの方が結果が見やすいかもしれない。

基本的な使い方としては、どちらの関数も、分析したいデータを数値行列として与えるだけで動作する。S との互換性のため、元データから分散共分散行列を計算し、それを使って主成分分析を行うのがデフォルトになっているが、それだと生データの絶対値の大きさに影響されてしまうので、\verb!princomp()!関数なら\verb!cor=TRUE!オプションをつけて、分散共分散行列でなく相関係数行列を使うようにすべきである。また、\verb!prcomp()!関数の場合は、\verb!scale=TRUE!オプションをつければ、各変数を標準化してから特異値分解してくれることになり、相関係数行列から出発するのとほぼ同じ結果が得られる。

\verb!princomp()!関数は素直に固有値と固有ベクトルを使って計算するため、変数の数がサンプルサイズより多いとエラーが出て計算できないが、\verb!prcomp()!関数は特異値分解によるため、変数の数がサンプルサイズより多くても計算できるという違いがある。

\verb!princomp()!も\verb!prcomp()!も主成分負荷量は出力しない。ただし、結果のオブジェクトを\verb!summary()!に与えると寄与率と累積寄与率は表示される。もう 1 つ表示されるのは standard deviation という値（変数名は\verb!sdev!）で固有値の平方根なので、その 2 乗をとれば各成分の固有値が得られる。

データ行列が\verb!X!だとすると、\verb!summary(princomp(X, cor=TRUE))$sdev^2!とすれば各主成分の固有値が得られる（\verb!summary(prcomp(X, scale=TRUE))$sdev^2!でも良い）。これは\verb!eigen(cor(X))$values!と同じである。

このとき主成分得点は、
\begin{screen}\begin{verbatim}princomp(X, cor=TRUE)$scores\end{verbatim}\end{screen}
または
\begin{screen}\begin{verbatim}prcomp(X, scales=TRUE)$x\end{verbatim}\end{screen}
で得られる\footnote{\url{https://blog.statsbeginner.net/entry/2014/07/27/121214}が参考になる。}。なお、\verb!princomp()!では分散などの計算で分母が\verb!N!だが\verb!prcomp()!では\verb!N-1!なので、微妙に結果は異なる。つまり、\verb!princomp()!では主成分得点の分散が固有値となっていて、\verb!prcomp()!では主成分得点の不偏分散が固有値となっているということ。\verb!prcomp()!の主成分得点を\verb!(N-1)/N!の平方根で割れば\verb!princomp()!が出す主成分得点と一致する。

Example 1 (PCA)

R の組み込みデータ\verb!swiss!は、1888 年頃のスイスのフランス語を話す 47 州について、標準化された出生力指標（変数名は\verb!Fertility!、Ig＝プリンストン研究（詳しくは、\verb!https://opr.princeton.edu/archive/pefp/indices.aspx!を参照されたい）の有配偶出生力指標で、既婚女性の出生率の生物学的上限と考えられるハテライトの出生率に対する比×100）、職業として農業に従事している男性の割合（同\verb!Agriculture!）、陸軍試験で最高ランクの評価を受けた被徴兵者の割合（同\verb!Examination!）、小学校より上の教育歴をもつ被徴兵者の割合（同\verb!Education!）、カソリック信者

の割合（同\verb!Catholic!）、乳児死亡率（同\verb!Infant.Mortality!）である。このデータを使って主成分分析を行い、これら 47 州のプロファイルを考えてみるコードを以下に示す。

```
\begin{screen}\small\begin{verbatim}
data(swiss)
spc <- princomp(swiss, cor=TRUE)
biplot(spc)
summary(spc)
summary(spc)$sdev^2
spc$loadings
\end{verbatim}\end{screen}
```

描かれるバイプロットは以下である。ここのコードでは表示されないが、各州の主成分得点を行列として欲しければ、\verb!spc$scores!で参照可能である。

$$\includegraphics[width=8cm]{swissprc.pdf}$$

```
\begin{screen}\scriptsize\begin{verbatim}
> summary(spc)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.7887865 1.0900955 0.9206573 0.66251693 0.45225403 0.34765292
Proportion of Variance 0.5332928 0.1980514 0.1412683 0.07315478 0.03408895 0.02014376
Cumulative Proportion  0.5332928 0.7313442 0.8726125 0.94576729 0.97985624 1.00000000
> summary(spc)$sdev^2
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
3.1997570 1.1883082 0.8476098 0.4389287 0.2045337 0.1208626
> spc$loadings

Loadings:
                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Fertility        -0.457  0.322  0.174  0.536  0.383 -0.473
Agriculture      -0.424 -0.412        -0.643  0.375 -0.309
Examination       0.510  0.125               0.814  0.224
Education         0.454  0.179 -0.532              -0.681
Catholic         -0.350  0.146 -0.807         0.183  0.402
Infant.Mortality -0.150  0.811  0.160 -0.527 -0.105

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
\end{verbatim}\end{screen}
```

この結果を表にまとめるときは以下のように負荷量の絶対値が小さいものは省略し、固有値と寄与率、累積寄与率を表示する（第 2 主成分までで十分かもしれないが、ここでは第 6 主成分まで書いた）。

$$\vbox{\small

\begin{tabular}{lrrrrrr}
\hline
 &\multispan{6}\hfill 主成分負荷量\hfill\cr
\cline{2-7}
変数 & 第 1 & 第 2 & 第 3 & 第 4 & 第 5 & 第 6\cr
\hline
有配偶出生力指数 & -0.457 & 0.322 & & 0.536 & 0.383 & -0.473\cr
男性農業従事割合 & -0.424 & -0.412 & & -0.643 & 0.375 & -0.309\cr
被徴兵者試験高成績割合 & 0.510 & & & & 0.814 & \cr
被徴兵者中等教育以上割合 & 0.454 & & -0.532 & & & -0.681\cr
カソリック信者割合 & -0.350 & & -0.807 & & & 0.402\cr
乳児死亡率 & & 0.811 & & -0.527 & & \cr
\hline
固有値 & 3.200 & 1.188 & 0.848 & 0.439 & 0.205 & 0.121\cr
寄与率 & 0.533 & 0.198 & 0.141 & 0.073 & 0.034 & 0.020\cr
累積寄与率 & 0.533 & 0.731 & 0.873 & 0.946 & 0.980 & 1.000\cr
\hline
\end{tabular}}$$

Example 2 (PCA)

\subsection{利用例 2}

実際に主成分分析を使って書かれた論文の中にはデータと解析結果が両方書かれているものがある。例として、Tokahoglu S (2012) Determination of trace elements in commonly consumed medicinal herbs by ICP-MS and multivariate analysis. {\it Food Chemistry\/}, 134: 2504-8.に掲載されている分析（著者は SPSS を使っている）を R で再現することを試みた。

結果のうち、次に示す Table 4 が主成分分析の結果である。

\begin{screen}\small
\vbox{\small
\begin{tabular}{lrrrr}
\noalign{Table 4. Varimax rotated loadings and communalities for herb samples}
\noalign{ (n = 30, only those larger than 0.1 are shown).}
\hline
Element &\multispan{4}\hfill Principal components\hfill & Communalities (h2)\cr
\cline{2-5}
& 1 & 2 & 3 & 4\cr
\hline
Cr & {\bf 0.917} & -0.182 & & 0.139 & 0.893\cr
Mn & 0.348 & & & {\bf 0.766} & 0.708\cr
Fe & {\bf 0.890} & & & 0.128 & 0.808\cr
Co & {\bf 0.946} & & & & 0.895\cr
Ni & {\bf 0.869} & 0.140 & 0.119 & 0.205 & 0.831\cr
Cu & 0.121& 0.108 & {\bf 0.811}& -0.324 & 0.789\cr
Zn & & -0.189 & {\bf 0.832} & 0.258 & 0.794\cr

Rb & & {\bf 0.725} & & {\bf 0.591} & 0.875\cr
Sr & & {\bf 0.898} & & -0.139 & 0.826\cr
Pb & {\bf 0.547} & {\bf -0.534} & 0.146 & & 0.606\cr\cr
Explained variance (\%) & 37.28 & 17.22 & 13.96 & 12.12 & 80.57\cr
\hline
\end{tabular}
}\end{screen}

論文には、バリマックス回転し、主成分負荷量の絶対値が 0.1 以上のものを表示したと書かれていた。分散共分散行列だとまったく違う結果になったので、相関係数行列を使っていると思われた。検出限界以下の扱いが不明であるが、\verb!princomp()!と\verb!prcomp()!では検出限界以下をタブ区切りテキストファイルには NA として入力したものを分析時に 0 に置換して処理した。青木繁伸教授の\verb!pca()!関数では自動的に欠損値を 1 つでも含むケースは除去される。これらのいずれも元論文と若干異なる結果であった。検出限界以下の値に対してペアワイズの除去をするために、\verb!cor()!関数のオプションで\verb!use="pairwise.complete.obs"!を使って相関係数行列を計算し、それを元に主成分分析を実行できる、\verb!psych!パッケージの\verb!principal()!を適用したところ、元論文と概ね合っている結果（微妙に違うが）が得られたので、おそらく元論文ではペアワイズの除去がなされたと考えられる。以上のコードを示しておく。

\begin{itembox}[l]{\url{https://minato.sip21c.org/advanced-statistics/MedHerbs.R}}
\scriptsize\begin{verbatim}
source("https://prs.ism.ac.jp/~nakama/AI/AI_UCS2.R") # 日本語 utf8 のため
par(family="Japan1GothicBBB") # pdf への日本語出力のため
windowsFonts(JP1=windowsFont("MS Gothic"),
 JP2=windowsFont("MS Mincho"), JP3=windowsFont("Meiryo"))
par(family="JP3") # Windows で画面でみるにはこちら。
Herbs <- read.delim("https://minato.sip21c.org/advanced-statistics/MedHerbs.txt")
row.names(Herbs) <- Herbs[, 1] # 最初の変数が薬草名なので行名にコピー
Herbs <- Herbs[, -1] # 薬草名を変数から削除
Herbsc <- Herbs # コピー
Herbs[sapply(Herbs, is.na)] <- 0 # このデータの NA は ND なので 0 を代入
# ただし ND の処理は難しい。検出限界以下はゼロではないので。
summary(res1 <- princomp(Herbs, cor=TRUE))
res1$sdev^2
res1$loadings
biplot(res1)
summary(res2 <- prcomp(Herbs, scale=TRUE, retx=TRUE))
res2$sdev^2
res2$rotation
biplot(res2)
# 違いは princomp では分母が N、prcomp では N-1 であること
# princomp では主成分得点の分散、prcomp では主成分得点の不偏分散が固有値
# 青木先生の関数 pca を読み込む
source("http://aoki2.si.gunma-u.ac.jp/R/src/pca.R", encoding="euc-jp")
res3 <- pca(Herbsc)
library(psych)
resx <- fa.parallel(Herbsc) # 出力する主成分数を決めるため

```
print(res3, npca=resx$ncomp)
print(res3, npca=4) # 強引に 4 つ出す
plot(res3)
# 手動でリスト単位の欠損値除去
Herbsc.omitNA <- subset(Herbsc, complete.cases(Herbsc))
summary(res1x <- princomp(Herbsc.omitNA, cor=TRUE))
res1x$loadings
# 合成得点を平均ゼロ、分散 1 に標準化するには、固有値の平方根で割ればいい
t(apply(res3$fs, 1, "/", sqrt(res3$eval))) # 主成分得点
t(apply(res1$scores, 1, "/", res1$sdev)) # 一致する
t(apply(res2$x, 1, "/", res2$sdev)) # 若干違う
#
# psych パッケージの principal()を使ってみる。主成分数を 4 にしたのは
# 元論文に合わせるため。それ以外の根拠はない。principal()はデフォルトで
# バリマックス回転する。
# principal()には相関係数行列しか与えられないので、主成分得点は出ない。
library(psych)
C1 <- cor(Herbsc, use="pairwise.complete.obs")
print(resp <- principal(C1, nfactors=4, n.obs=length(Herbsc[, 1]))) # 元論文で主成分が 4 つなので
\end{verbatim}\end{itembox}
```

まず、\verb!princomp()!の結果を示す。以下の枠内の通り、絶対値でみると、第 1 主成分負荷量が大きい元素は Cr、Fe、Co、Ni、第 2 主成分負荷量が大きい元素が Rb と Sr、第 3 主成分負荷量が大きい元素が Cu と Zn、第 4 主成分負荷量が大きい元素が Mn と Zn となっており、微妙に違っているが概ね論文に掲載されている表と同じ傾向になっていることがわかる（負荷量の値自体はまるで違うが）。第 4 主成分までの寄与率も 80.56\%であり、元論文の表とほぼ同じである。

```
\begin{screen}\small\begin{verbatim}
Importance of components:
                          Comp.1    Comp.2    Comp.3
Standard deviation     2.0020096 1.3577846 1.1389361
Proportion of Variance 0.4008042 0.1843579 0.1297175
Cumulative Proportion  0.4008042 0.5851621 0.7148797
                          Comp.4     Comp.5     Comp.6
Standard deviation     0.95244412 0.88821102 0.73826604
Proportion of Variance 0.09071498 0.07889188 0.05450367
Cumulative Proportion  0.80559464 0.88448652 0.93899019
                          Comp.7     Comp.8     Comp.9
Standard deviation     0.52824069 0.39781011 0.34428285
Proportion of Variance 0.02790382 0.01582529 0.01185307
Cumulative Proportion  0.96689402 0.98271930 0.99457237
                          Comp.10
Standard deviation     0.232972709
Proportion of Variance 0.005427628
Cumulative Proportion  1.000000000

> res1$sdev^2
```

```
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
4.00804224 1.84357898 1.29717535 0.90714981 0.78891881
    Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
0.54503675 0.27903822 0.15825288 0.11853068 0.05427628
\end{verbatim}\end{screen}

\begin{screen}\small\begin{verbatim}
> res1$loadings

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
Cr  0.461        -0.136  0.107  0.125 -0.110  0.113  0.276
Mn  0.261 -0.160 -0.182 -0.615 -0.520  0.334  0.153 -0.251
Fe  0.445               0.137  0.105 -0.295 -0.468 -0.631
Co  0.461               0.216         0.309  0.137 -0.130
Ni  0.444 -0.135               0.237  0.314         0.409
Cu         0.198  0.722  0.134 -0.392  0.287 -0.373  0.125
Zn  0.131  0.256  0.535 -0.512  0.416 -0.222  0.349 -0.145
Rb        -0.630  0.109 -0.359        -0.291 -0.405  0.365
Sr        -0.555  0.337  0.361 -0.257 -0.235  0.533 -0.166
Pb  0.291  0.369 -0.106        -0.493 -0.569         0.284
   Comp.9 Comp.10
Cr  0.766 -0.224
Mn        -0.146
Fe -0.112 -0.228
Co         0.773
Ni -0.530 -0.408
Cu  0.154
Zn
Rb         0.266
Sr        -0.123
Pb -0.298  0.151

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
SS loadings       1.0    1.0    1.0    1.0    1.0    1.0
Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1
Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6
               Comp.7 Comp.8 Comp.9 Comp.10
SS loadings       1.0    1.0    1.0     1.0
Proportion Var    0.1    0.1    0.1     0.1
Cumulative Var    0.7    0.8    0.9     1.0
\end{verbatim}\end{screen}

$$\includegraphics[width=10cm]{MedHerbsPCA1.pdf}$$

\verb!prcomp()!の結果は以下の通りであり、回転後の負荷量を示すはずの\verb!res2$rotation!をみても、概ね\verb!princomp()!と同じ結果になった。
```

\begin{screen}\small\begin{verbatim}
> summary(res2 <- prcomp(Herbs, scale=TRUE, retx=TRUE))
Importance of components:
                          PC1   PC2   PC3   PC4   PC5
Standard deviation   2.0020 1.3578 1.1389 0.95244 0.88821
Proportion of Variance 0.4008 0.1844 0.1297 0.09071 0.07889
Cumulative Proportion 0.4008 0.5852 0.7149 0.80559 0.88449
                          PC6   PC7   PC8   PC9   PC10
Standard deviation   0.7383 0.5282 0.39781 0.34428 0.23297
Proportion of Variance 0.0545 0.0279 0.01583 0.01185 0.00543
Cumulative Proportion 0.9390 0.9669 0.98272 0.99457 1.00000

> res2$sdev^2
 [1] 4.00804224 1.84357898 1.29717535 0.90714981 0.78891881
 [6] 0.54503675 0.27903822 0.15825288 0.11853068 0.05427628
\end{verbatim}\end{screen}

\begin{screen}\small\begin{verbatim}
> res2$rotation
          PC1   PC2   PC3   PC4
Cr -0.46083242 0.06364328 -0.1361550921 0.107486068
Mn -0.26136598 -0.16035696 -0.1823044797 -0.615253775
Fe -0.44463160 -0.05663739 0.0001361183 0.136786984
Co -0.46117884 -0.05662732 -0.0280148314 0.216279962
Ni -0.44365438 -0.13532975 0.0532524244 0.071228437
Cu -0.07913501 0.19762575 0.7220481058 0.133815362
Zn -0.13138608 0.25554318 0.5354282736 -0.511640878
Rb -0.06136051 -0.63034651 0.1085422052 -0.359462244
Sr 0.01080989 -0.55467009 0.3369436981 0.361134433
Pb -0.29103660 0.36898876 -0.1060982013 0.005850113
          PC5   PC6   PC7   PC8   PC9
Cr -0.12549872 0.1096002 -0.11309434 0.2757762 -0.76584041
Mn 0.51993276 -0.3340441 -0.15330073 -0.2507013 -0.06783438
Fe -0.10536233 0.2951697 0.46795224 -0.6313140 0.11197090
Co -0.08484872 -0.3090026 -0.13747358 -0.1304947 0.02462569
Ni -0.23723424 -0.3138834 -0.08751828 0.4093297 0.52995697
Cu 0.39195326 -0.2872190 0.37327270 0.1252577 -0.15405082
Zn -0.41622196 0.2222648 -0.34873144 -0.1448879 0.01531580
Rb -0.06080207 0.2908507 0.40496476 0.3648576 -0.04006951
Sr 0.25723944 0.2350568 -0.53290082 -0.1657695 -0.01525029
Pb 0.49254674 0.5687966 -0.09615068 0.2840365 0.29846994
         PC10
Cr 0.22444386
Mn 0.14605493
Fe 0.22821852
Co -0.77287087
Ni 0.40752896
Cu 0.02336532
Zn -0.05909749

Rb -0.26625092
Sr 0.12310930
Pb -0.15071856
\end{verbatim}\end{screen}

$$\includegraphics[width=10cm]{MedHerbsPCA2.pdf}$$

一方、測定限界以下を 0 にせずに欠損値としてリスト単位で除去する（＝1 つでも欠損値があれば、その薬草データごと除去する）青木先生の\verb!pca()!の結果は、\verb!fa.pararrel()!では適切な主成分数が 2 となったが、より多くの主成分について結果を表示しても Contribution が増えるだけで、第 1 主成分や第 2 主成分についての負荷量や寄与率は変わらないので、4 つの主成分について負荷量と寄与率を下表に示すこの表示だと負荷量の絶対値は元論文の値に近づくが、第 2 主成分が Rb と Sr ではなく Mn と Rb になり、第 3 主成分が Cu と Zn でなく Cu と Sr になり、第 4 主成分が Mn と Rb でなく Zn と Rb になるという大きな違いが出てしまうので、おそらく元論文の欠損値処理はリスト単位の除去ではない。なお、主成分には名前をつける必要はないので、ここでも\verb!PC1!などとしている。

	PC1	PC2	PC3	PC4	Contribution
Cr	-0.919	-0.168	0.120	-0.076	0.893
Mn	-0.504	0.628	0.041	-0.015	0.650
Fe	-0.890	-0.164	0.025	0.047	0.822
Co	-0.921	-0.053	-0.139	-0.122	0.885
Ni	-0.900	0.030	-0.083	0.115	0.832
Cu	0.029	-0.328	-0.808	0.242	0.820
Zn	-0.050	-0.322	-0.060	0.883	0.889
Rb	-0.311	0.770	0.080	0.440	0.890
Sr	-0.108	0.298	-0.824	-0.308	0.875
Pb	-0.477	-0.545	0.196	-0.204	0.605
Eigenvalue	3.889	1.644	1.423	1.204	
Contribution	0.389	0.164	0.142	0.120	
Cum.contrib.	0.389	0.553	0.696	0.816	

手動でリスト単位の除去を行い、\verb!princomp()!を使って計算した結果も、負荷量の絶対値は測定限界以下にゼロを入れた場合と似ているが、主成分ごとに負荷量の高い元素をみると、青木先生の\verb!pca()!を使った場合と同じく、元論文のパタンと大きく食い違っているので、やはりリスト単位の除去ではないと考えられる。

そこで、測定限界以下を欠損値としてペア単位の除去（変数 2 つずつの組合せごとに、どちらかが欠損ならば、その 2 つの変数間の相関係数の計算からのみ除去）をして相関係数行列を求め、それを入力にした\verb!psych!パッケージの\verb!principal()!関数の結果は以下のように得られた。これはほぼ論文に示されている結果と一致しているので（微妙に違うが）、同じ方法と考えて良いだろう。

\begin{screen}\small\begin{verbatim}
Principal Components Analysis
Call: principal(r = C1, nfactors = 4, n.obs = length(Herbsc[, 1]))
Standardized loadings (pattern matrix) based upon correlation matrix
   RC1   RC2   RC3   RC4   h2   u2 com
Cr 0.92 -0.14 0.04 0.11 0.89 0.11 1.1

```
Mn 0.40 -0.08 -0.02  0.64 0.58 0.42 1.7
Fe 0.89  0.08  0.09  0.10 0.82 0.18 1.1
Co 0.92  0.06  0.03  0.11 0.86 0.14 1.0
Ni 0.85  0.14  0.18  0.27 0.84 0.16 1.4
Cu 0.11  0.21  0.76 -0.41 0.80 0.20 1.8
Zn 0.07 -0.23  0.87  0.21 0.85 0.15 1.3
Rb 0.03  0.64 -0.04  0.65 0.84 0.16 2.0
Sr 0.02  0.94 -0.05 -0.07 0.89 0.11 1.0
Pb 0.60 -0.35 -0.09 -0.33 0.60 0.40 2.3

                       RC1  RC2  RC3  RC4
SS loadings           3.75 1.57 1.38 1.28
Proportion Var        0.37 0.16 0.14 0.13
Cumulative Var        0.37 0.53 0.67 0.80
Proportion Explained  0.47 0.20 0.17 0.16
Cumulative Proportion 0.47 0.67 0.84 1.00

Mean item complexity =  1.5
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.08
 with the empirical chi square  17.72  with prob <  0.088

Fit based upon off diagonal values = 0.95
\end{verbatim}\end{screen}
```

この表の\verb!h2!は、\verb!pca()!の\verb!Contribution!と同じで、いわゆる共通性(communality)を示すものとして元論文に掲載されている値になる。これを元論文と同じく絶対値が 0.1 以上のものだけ残して表にしたものを表\ref{tbl:pcelements}に示す。

```
\begin{table}[ht]
\caption{薬草中の元素含有量についての主成分分析結果 \label{tbl:pcelements}}
\begin{tabular}{lccccc}
\hline
 &\multispan{4}\hfill Loadings \hfill\cr
\cline{2-5}
Elements & PC1 & PC2 & PC3 & PC4 & Communality ($h^2$)\cr
\hline
Cr & 0.92 & $-0.14$ &   &  0.11 & 0.89\cr
Mn & 0.40 &        &      &  0.64 & 0.58\cr
Fe & 0.89 &        &      &       & 0.82\cr
Co & 0.92 &        &      &       & 0.86\cr
Ni & 0.85 & 0.14   & 0.18 &  0.27 & 0.84\cr
Cu & 0.11 & 0.21 & 0.76 &  $-0.41$ & 0.80\cr
Zn &      & $-0.23$ & 0.87 & 0.21 & 0.85\cr
Rb &      & 0.64  &      &  0.65   & 0.84\cr
Sr &      & 0.94    &      &       & 0.89\cr
Pb & 0.60 & $-0.35$ &      & $-0.33$ & 0.60\cr
```

\hline
Statistics & PC1 & PC2 & PC3 & PC4 & \cr
Eigenvalues & 3.75 & 1.57 & 1.38 & 1.28 & \cr
Proportion Variance & 0.37 & 0.16 & 0.14 & 0.13 & \cr
Cumulative Variance & 0.37 & 0.53 & 0.67 & 0.80 & \cr
\hline
\end{tabular}
\end{table}

Basic usage of factor analysis

I recommend to read the document titled “A Primer on Factor Analysis in Research using Reproducible R Software”[7].

\begin{description}
\item[入力データ] ある程度のサンプルサイズと大きな変数をもつ数値行列で、通常、サンプルサイズは{\bf 300 より多い。} {\bf 変数数に対する対象者の人数の比}は、通常、{\bf 2:1 から 10:1}の範囲をとる。原則として変数は正規分布に従うべきでだし、外れ値は含まない方がよい。他の変数と関連のない変数は分析に含めるべきではない。お互いに相関係数 1.0 の変数は含めることができない。どちらかを除外するか、適切であれば両者の和をとって合成変数として用いることは可能である。
\item[出力] (1) 因子負荷量は、各変数がその元になる潜在因子と相関している程度を意味する（その際、さまざまな回転が用いられる\footnote{最初の因子負荷量は、第一因子への負荷を最大にするように計算されるので、たいていの変数が１つ以上の因子に対して高い負荷量をもってしまい、因子の解釈が難しくなる。そこで、適切な{\bf 回転}をすると、この問題が解決することが多い。}）、 (2) 因子得点は、通常、各個人の応答と因子負荷量の積の和で（ただし複数の計算法があり、どの方法が最適かについて統一見解はない）、各個人の特性がどの程度その因子によって説明されるかを示す。
\item[回転] 回転の方法は２つに大別される。直交回転は、因子間の独立性を保ったまま因子ベクトルを回転させるが、斜交回転では因子間に相関が出てもいいことにしている。因子が理論的に相互依存を許してもいいときに、後者を考えるべきである。前者には最もよく使われていて単純なバリマックス回転が含まれる。バリマックス回転は、因子ごとの分散を最大化する。後者にはプロマックス回転やオブリミン回転が含まれる。
\item[因子分析のための道具] スクリープロット、バートレットの球面性検定、カイザー・マイヤー・オルキンのサンプリング適切性基準、平行分析(Parallel Analysis)が便利。因子数がうまく決定できたら、各因子に含まれる変数が単一軸の加法的スコアになっているかどうかをチェックするために、クロンバックのα係数を計算する（通常、それらの因子の和が信頼できるスコアであるためには、クロンバックのαが 0.7 より大きくなければいけない）。
\end{description}

推定された因子を解釈する際には、{\bf 因子に適切な名前（意味）をつけることが必要}である。因子がうまく推定できたと判定するには、因子負荷量が高い変数が少なくとも３つあるべきである。もし１つか２つしか因子負荷量が高い変数がないときは、因子数が多すぎるか、元の変数間に多重共線性が存在する可能性がある。

7 https://rpubs.com/Geesaale/1064415

The basic model of factor analysis

\section{因子分析の基本モデル}

300 人で変数 10 個（$X_1, X_2, ..., X_{10}$）の場合を考えよう。これら 10 個の変数の背後に、もし 2 個の潜在因子（F_1とF_2）があるとしたら、各変数は、これらの因子によって次のように説明される。

$$X_1 = \beta_{1\cdot 1} F_1 + \beta_{2\cdot 1} F_2 + \epsilon_1 $$
$$X_2 = \beta_{1\cdot 2} F_1 + \beta_{2\cdot 2} F_2 + \epsilon_2 $$
$$\vdots$$
$$X_{10} = \beta_{1\cdot 10} F_1 + \beta_{2\cdot 10} F_2 + \epsilon_{10} $$

ここで、βは、各変数と潜在因子との相関を意味し、これを{\bf 因子負荷量 (Factor loadings)}と呼ぶ。ϵは誤差分散を意味する。言い換えると、推定された因子では説明できなかった{\bf 独自性 (uniqueness)}でもある。なお、独自性を 1 から引いたものを共通性(communality)という。後述する\verb!rela!パッケージの\verb!!関数では、共通性が出力される。しかし、潜在因子F_1とF_2は測定された値ではない。だから、我々は、主因子法、最小残差法、最尤法などの様々な方法で、反復計算させながら推定しなくてはいけない\footnote{主成分分析では、各主成分は、測定された変数の線形結合として定式化されるので、反復推定は必要ない。}。

回転する前は、因子F_1とF_2は独立と仮定されている。いま、n番目（n は区間[1, 300]の整数）の人のi番目の変数の値を$X_i(n)$と書くと、その人の因子得点（ここでは$FS_1(n)$と$FS_2(n)$）は、次のように得られる（ただし、これは最も単純な方法である。因子得点として提案されている指標値は、この他にもいくつかある）。計算に使う変数は、βの絶対値が十分大きい（通常、0.3 とか 0.4、あるいは 0.5 以上とする）ものに限るのが普通。

$$FS_1(n) = \sum^{10}_{i=1} \beta_{1\cdot i} X_i(n)$$
$$FS_2(n) = \sum^{10}_{i=1} \beta_{2\cdot i} X_i(n)$$

How many factors should be estimated?

この問題には以下のようにいくつかの基準が提案されているが、100\%これが良いという検定法などは存在しない。

\begin{description}
\item[スクリープロットを描く] 最初に可能な限り多くの因子を仮定して因子分析を行い、各因子によって説明される分散を代表するものとしての固有値（あるいは同じ意味で因子負荷量の二乗和）を、大きい順に線でつないだ折れ線グラフがスクリープロットである。折れ線が急に激しく落ち込む変数があれば、その直前が適切な因子数と考えられる。
\item[パラレル分析をする] 実際のスクリープロットを、ランダムにリサンプルしたデータから計算したスクリープロットと比較する。２つのプロットが交差する点が適切な因子数であると考える。
\item[固有値が 1 を超えている間] 固有値が 1 を超えている間は、変数 1 つよりも情報量が多いと考えられるので。
\end{description}

Checking the sampling adequacy of factor analysis

\section{因子分析の適切性をチェックする}

因子分析の適切性をチェックするための方法がいくつかある。

\begin{description}
\item[サンプルサイズの適切性の基準] サンプルサイズは 50 では非常に乏しい(very poor)。100 でも乏しい(poor)。200 ならまあまあ(fair)、300 なら十分(good)、500 なら非常に良い(very good)。1,000 を超えたら極めて優れている(excellent)といえる(Comfrey and Lee, 1992, p.217)。
\item[KMO と MSA] KMO とは、Kaiser-Meyer-Olkin が提唱した因子分析全体についてのサンプリング適切性基準であり、MSA とは Measures of Sampling Adequacy の頭語で、それぞれの変数についての個別のサンプリング適切性基準である。データセットの中に、十分な数の因子が存在するかどうかを示す指標値である。技術的には、変数間の相関係数の偏相関係数に対する比を計算する。もし偏相関係数が生の相関係数と同じような値なら、それらの変数は互いに分散をあまり共有していないことを意味する。
KMO の範囲は 0.0 から 1.0 で、0.5 以上が望ましい\footnote{Kaiser (1974)の提案によれば、0.5 未満では不適切、0.5 以上 0.6 未満は悲惨なレベル(miserable)、0.6 以上 0.7 未満は良くも悪くもなく(mediocre)、0.7 以上 0.8 未満は並(middling)、0.8 以上 0.9 未満は賞賛に値し(meritorious)、0.9 以上なら極めて優れている(marvelous)。}。また、MSA が 0.5 未満の変数は、その変数がどの因子グループにも属していないことを示すので、因子分析から除くべきである。
\begin{screen}\small
群馬大学の青木繁伸教授は、\url{http://aoki2.si.gunma-u.ac.jp/R/kmo.html}で、KMO と MSA を計算するための次の関数定義を公表している\footnote{source("http://aoki2.si.gunma-u.ac.jp/R/src/kmo.R", encoding="euc-jp")で使えるようになる。}。

\begin{verbatim}
kmo <- function(x)
{
	x <- subset(x, complete.cases(x)) #欠損値除去
	r <- cor(x) # 相関係数行列を r に付値
	r2 <- r^2 # 相関係数行列の各要素を 2 乗した値を r2 に付値
	i <- solve(r) # 相関係数行列 r の逆行列を求めて i に付値
	d <- diag(i) # 逆行列 i の対角成分を d に付値
	p2 <- (-i/sqrt(outer(d, d)))^2 # 偏相関係数の 2 乗を計算し p2 に付値
	diag(r2) <- diag(p2) <- 0 # r2 と p2 の対角成分を 0 にする
	KMO <- sum(r2)/(sum(r2)+sum(p2))
	MSA <- colSums(r2)/(colSums(r2)+colSums(p2))
	return(list(KMO=KMO, MSA=MSA))
}
\end{verbatim}\end{screen}
\item[バートレットの球面性検定] 変数間の相関が偶然期待されるより大きいという仮説を検定する。技術的には行列が単位行列であるかどうかを検定する。p 値が有意である場合、対角以外のすべての相関がゼロであるという帰無仮説が棄却される。
\begin{screen}\small

バートレットの球面性検定についても、群馬大学の青木繁伸教授が\url{http://aoki2.si.gunma-u.ac.jp/R/Bartlett.sphericity.test.html}で次の関数定義を公表している\footnote{source("http://aoki2.si.gunma-u.ac.jp/R/src/Bartlett.sphericity.test.R", encoding="euc-jp")で使えるようになる。}。

\begin{verbatim}
Bartlett.sphericity.test <- function(x)
{
    method <- "Bartlett's test of sphericity"
    data.name <- deparse(substitute(x))
    x <- subset(x, complete.cases(x)) # 欠損値除去
    n <- nrow(x)
    p <- ncol(x)
    chisq <- (1-n+(2*p+5)/6)*log(det(cor(x)))
    df <- p*(p-1)/2
    p.value <- pchisq(chisq, df, lower.tail=FALSE)
    names(chisq) <- "X-squared"
    names(df) <- "df"
    return(structure(list(statistic=chisq, parameter=df,
        p.value=p.value, method=method, data.name=data.name),
        class="htest"))
}
\end{verbatim}\end{screen}
\end{description}

The functions to conduct factor analysis in R

\begin{screen}
以下に説明するように、追加パッケージとして\verb!psych!、\verb!sem!を用いるので、
\begin{verbatim}
install.packages("psych", dep=TRUE)
install.packages("sem", dep=TRUE)
\end{verbatim}
として、予めインストールされたい。
\end{screen}

\begin{description}
\item[factanal] この関数は標準でインストールされる。因子負荷量を計算するのに最尤法を用いる。推定すべき因子数は明示的に指定せねばならない。バリマックス回転とプロマックス回転が可能である。入力データは行列またはデータフレーム。
\item[fa] この関数は{\bf psych}パッケージに含まれている。\verb!fm=!オプションで因子負荷量の計算方法を指定できる（\verb!"minres"!で最小残差法、\verb!"ml"!で最尤法、\verb!"pa"!で主因子法）。推定する因子数は\verb!nfactors=!オプションで指定せねばならない。\verb!rotate=!オプションでさまざまな回転方法を指定できる（\verb!"none"!、\verb!"varimax"!、\verb!"quartimax"!、\verb!"bentlerT"!、\verb!"geominT"!、\verb!"oblimin"!、\verb!"simplimax"!、\verb!"bentlerQ"!、\verb!"geominQ"!、\verb!"cluster"!が可能）。

\item[alpha] この関数は{\bf psych}\index{psych}パッケージに含まれている。クロンバックのα係数\index{くろんばっくのあるふぁけいすう@クロンバックのα係数}を計算する。
\item[cortest.bartlett] この関数も{\bf psych}パッケージに含まれている。バートレットの球面性検定を実行する。
\item[fa.parallel] この関数も{\bf psych}パッケージに含まれている。パラレル分析を実行し、返り値として、\verb!$nfact!に推定すべき適切な因子数を返す。
\item[sem] 確証的因子分析(confirmatory factor analysis; CFA)には、{\bf sem}パッケージを用いることができる。もちろん sem は構造方程式モデリングのパッケージであり、CFA 以上のことができる。詳しくは次章で触れる。
\end{description}

Example of factor analysis using ecopoint data

データを使って実例を示そう。\url{https://minato.sip21c.org/advanced-statistics/ecopx.txt}は、\url{https://minato.sip21c.org/humeco/ecopoint.html}に示したエコポイントチェック（図\ref{fig:ecopointentryform}）への回答\footnote{かつて web サイトで cgi を使ってくださった、大勢の匿名の皆様に感謝申し上げる。}を適当に加工した、タブ区切りテキストデータである。

\begin{figure}[ht]
\begin{center}
\includegraphics[width=12cm]{ecopointform.pdf}
\end{center}
\caption{エコポイントチェックの入力フォーム\label{fig:ecopointentryform}}
\end{figure}

エコポイントとは、高月紘（編著）『自分の暮らしがわかるエコロジー・テスト：環境問題は生活のエコ度チェックから』講談社ブルーバックスに提示されている、どの程度「環境に優しい」暮らしをしているかを示す尺度である。評点の重み付けには、環境研のコンパラティブ・リスクアセスメントの結果を使っており、ある程度妥当な評価尺度と考えられる。この質問紙を web から回答できるように cgi 化し、不特定多数から回答を得た。

エコポイント総合点（変数名\verb!EP!）が 99.8 点満点（計算上の丸め誤差のため）、温暖化問題エコポイント（\verb!GW!）が 14.3 点満点、廃棄物問題エコポイント（\verb!Waste!）が 24.6 点満点、水質汚染問題エコポイント（\verb!Water!）が 15.6 点満点、大気汚染問題エコポイント（\verb!Air!）が 21.1 点満点、有害化学物質問題エコポイント（\verb!Chem!）が 24.2 点満点で表示される。

書籍によれば、若い層を中心とした対象者 356 人の平均が 42.6、環境問題の講演を聞きに来た人たち 182 人の平均が 48.1 という指標である。著者は、環境にやさしい人としては 60 点以上必要で、30 点以下だったら環境面ではかなり問題のあるライフスタイルとしており、低い場合は、どの行動パタンにとくに問題があるのかをチェックすることが薦められている。

Calculation of Cronbach’s alpha coefficient

このデータを\verb!eco!というデータフレームに読み込み、まずは計算された５つのエコポイント得点について、クロンバックのα係数を計算してみる。ここで、１つのデータに対して多くの解析をするので、専用のディレクトリを作ると良い。ここでは\verb!c:/work/lecture/kobe/advstat2/!というディレクトリを作り、ここに\url{https://minato.sip21c.org/advanced-statistics/ecopx.txt}をダウンロードしてから、RStudio を起動し、\verb!File!の\verb!New project!から、\verb!Existing Directory!として\verb!e:/work/lecture/kobe/advstat2/!を選択した。こうすると、RStudio 終了時に、自動的にそのときの環境がこのディレクトリの\verb!advstat2.Rproj!というファイルに保存される。次回からは、このファイルをダブルクリックするだけで自動的に RStudio が起動し、前回最後に操作していたときの状態が復元される。

次の枠内に示すコードをこのディレクトリにダウンロードして右下ペインから選べば左上にスクリプトエディタ画面が開くが、\verb!File!の\verb!New File!の\verb!R script!で白紙のスクリプトエディタ画面を開いて、ブラウザ等で開いたコードをコピーアンドペーストしても良い。スクリプトエディタウィンドウ右上の\verb!Source!というボタンをクリックすると、自動的に５つのエコポイント得点それぞれについて、クロンバックのα係数と 95\%信頼区間が計算される。ちなみに、\verb!alpha()!関数の返り値には推定値（\verb!raw_alpha!）と漸近標準誤差（\verb!ase!）は含まれているが、普通に実行すると表示される信頼区間の上限と下限の計算は、\verb!print.psych()!に含まれている。以下のコードでは、\verb!GAC()!という関数を定義して、\verb!alpha()!の結果のうち推定値と 95\%信頼区間の下限と上限だけを返すようにした。1.96 は言うまでもなく正規分布の 97.5\%点、即ち\verb!qnorm(0.975)!であり、90\%信頼区間が欲しいときは 1.96 の部分を\verb!qnorm(0.95)!とすれば良い。

\begin{itembox}[l]{\url{https://minato.sip21c.org/advanced-statistics/ecopxc.R}}
\small\begin{verbatim}
eco <- read.delim("ecopx.txt")
# 前処理
eco$NAGE <- factor(eco$AGE+1,
  labels=c("10-19","20-29","30-39","40-49","50-59","60-69","70-"))
eco$NSEX <- factor(eco$SEX+1, labels=c("M","F"))
warming <- eco[, c("FAMSIZE","Q05","Q07","Q08","Q11","Q24")]
waste <- eco[, c("FAMSIZE","Q01","Q02","Q03","Q04","Q06")]
water <- eco[, c("FAMSIZE","Q13","Q14","Q16","Q17","Q20")]
air <- eco[, c("FAMSIZE","Q09","Q10","Q12","Q23","Q25")]
chem <- eco[, c("FAMSIZE","Q15","Q18","Q19","Q21","Q22")]
ecopoint <- eco[, c("FAMSIZE","Q05","Q07","Q08","Q11","Q24",
 "Q01","Q02","Q03","Q04","Q06","Q13","Q14","Q16","Q17","Q20",
 "Q09","Q10","Q12","Q23","Q25","Q15","Q18","Q19","Q21","Q22")]
library(psych)
# α と信頼区間を得るための関数定義
GAC <- function(Z) { # Get alpha /w 95 percent confidence intervals
 ZA <- alpha(Z)
 Raw <- ZA$total$raw_alpha
 Ase <- ZA$total$ase
 return(c(Raw-1.96*Ase, Raw, Raw+1.96*Ase))
}
all <- cbind(GAC(warming[,-1]), GAC(waste[,-1]), GAC(water[,-1]),
          GAC(air[,-1]), GAC(chem[,-1]), GAC(ecopoint[,-1]))
```
```

```
print(all)
\end{verbatim}\end{itembox}
```

結果を見ると、クロンバックのα係数は、全項目を使ったエコポイントとしては 0.84 [0.81-0.88]と十分に高いが、温暖化領域 0.41 [0.28-0.53]、廃棄物領域 0.61 [0.52-0.71]、水領域 0.69 [0.60-0.78]、大気領域 0.43 [0.32-0.55]、化学物質領域 0.66 [0.57-0.75]であり、各領域は 0.7 以上という基準に達していない。おそらく多様な回答者に対して設問が微妙な答えにくさを含んでいるためと、法制の影響などもあるものと思われるが、尺度としての信頼性は十分でない。そこで、単身者と 2 人以上で生活している人で構造が違う可能性を考え、それぞれサブセットを作って分析してみたが、大差なかった（図\ref{fig:ecopxc}）。

```
\begin{itembox}[l]{ecopxc.R の続き}\small\begin{verbatim}
# for single household
single <- cbind(
GAC(subset(warming, FAMSIZE==1)[,-1]), GAC(subset(waste, FAMSIZE==1)[,-1]),
GAC(subset(water, FAMSIZE==1)[,-1]), GAC(subset(air, FAMSIZE==1)[,-1]),
GAC(subset(chem, FAMSIZE==1)[,-1]), GAC(subset(ecopoint, FAMSIZE==1)[,-1]))
print(single)
# for other household
others <- cbind(
GAC(subset(warming, FAMSIZE>1)[,-1]), GAC(subset(waste, FAMSIZE>1)[,-1]),
GAC(subset(water, FAMSIZE>1)[,-1]), GAC(subset(air, FAMSIZE>1)[,-1]),
GAC(subset(chem, FAMSIZE>1)[,-1]), GAC(subset(ecopoint, FAMSIZE>1)[,-1]))
print(others)
# まとめる
MX <- rbind(all[2,], single[2,], others[2,])
colnames(MX) <- c("温暖化","廃棄物","水","大気","化学物質","総合")
rownames(MX) <- c("全体","単独世帯","他の世帯")
UX <- rbind(all[3,], single[3,], others[3,]) # 95\%信頼区間の上限
# cairo_pdf("ecopxc.pdf")
# source("https://prs.ism.ac.jp/~nakama/AI/AI_UCS2.R")
# par(family="Japan1Gothic", las=1)
par(family="sans", las=1) # Windows で画面表示ならこれで良い
ii <- barplot(MX, beside=TRUE, ylim=c(0,1), col=1:3)
arrows(ii, as.vector(MX), ii, as.vector(UX), angle=90, length=0.1)
legend("topleft", legend=rownames(MX), fill=1:3, cex=0.6)
# dev.off()
\end{verbatim}\end{itembox}
```

```
\begin{figure}[ht]
\begin{center}
\includegraphics[width=10cm]{ecopxc.pdf}
\end{center}
\caption{クロンバックの$\alpha$係数と 95\%信頼区間の上限、世帯のタイプ別\label{fig:ecopxc}}
\end{figure}
```

Try to conduct the exploratory factor analysis

各質問項目の選択肢に与えたスコアの重みはかつて環境省で行われたコンパラティブ・リスクアセスメント(CRA)の結果によるので、それは生かすことにして\footnote{ただし、本当にこのスコアで良いのか、むしろ、元々のスコアのまま標準化した方が良いのではないかという問題はあるので、その辺りは今後丁寧に検討すべきである。}、しかし各下位尺度のクロンバックのα係数が低いので、おそらく回答者の違いや時代の違いにより、因子構造が想定と合っていないのだと判断し、\verb!Q01!から\verb!Q25!のデータを探索的因子分析してみる。

\begin{itembox}[l]{\url{https://minato.sip21c.org/advanced-statistics/ecofactor.R}}
\small\begin{verbatim}
eco.raw <- eco[,4:28]
source("http://aoki2.si.gunma-u.ac.jp/R/src/kmo.R", encoding="euc-jp")
kmo(eco.raw)
library(psych)
cortest.bartlett(eco.raw)
print(res1 <- fa.parallel(eco.raw))
print(res2 <- fa(eco.raw, fm="minres", nfactors=res1$nfact,
 rotate="quartimax"))
res2$loadings
\end{verbatim}\end{itembox}

群馬大学青木繁伸教授の関数で KMO や MSA を出すと概ね 0.8 以上あるので十分である。\verb!cortest.bartlett()!の結果、p 値はほぼ 0 であり、回答に相関がないという帰無仮説が棄却されるので因子分析に適したデータといえる。\verb!fa.parallel()!の結果、``Parallel analysis suggests that the number of factors = 5 and the number of components = 4"と表示されるので因子数は想定通り 5 で良いと考えられる（図\ref{fig:pascreeplot}）。

\begin{figure}[h]
\begin{center}
\includegraphics[width=10cm]{pascreeplot.pdf}
\end{center}
\caption{パラレル分析とスクリープロットによる因子数探索\label{fig:pascreeplot}}
\end{figure}

そこでクォーティマックス回転して因子分析をすると以下が得られる。

\begin{screen}\small\begin{verbatim}
    MR2   MR1   MR3   MR5   MR4
Q01 0.176 0.166   0.625
Q02 0.293 0.164 0.287   0.269
Q03 0.248 0.199 0.169   0.670
Q04 0.158 0.296 0.465 0.208 0.198
Q05 0.244 0.261 0.161 0.516 0.189
Q06 0.296 0.145   0.625 0.100
Q07 0.241 0.193   0.350
Q08   0.322   0.277
Q09 0.355   0.422 -0.200

```
Q10  0.365        0.155  0.375
Q11 -0.176  0.137  0.581
Q12  0.316  0.336  0.360  0.145
Q13  0.138  0.181  0.527        0.122
Q14  0.216  0.246  0.451  0.212  0.237
Q15  0.502  0.323  0.457
Q16  0.385  0.300  0.554 -0.108
Q17  0.462        0.202  0.130  0.323
Q18  0.687  0.156              0.137
Q19  0.462  0.106        0.320  0.296
Q20  0.632  0.269  0.130
Q21  0.650              0.164
Q22  0.202  0.789  0.174
Q23        0.930
Q24        0.428  0.236  0.191
Q25  0.234  0.142        0.315  0.278

                MR2   MR1   MR3   MR5   MR4
SS loadings    3.102 2.560 2.129 1.535 1.516
Proportion Var 0.124 0.102 0.085 0.061 0.061
Cumulative Var 0.124 0.226 0.312 0.373 0.434
\end{verbatim}\end{screen}
```

第 5 因子まで入れても分散の 43.4\%しか説明できないし、どの因子ともあまり関係していない変数が多々ある。これは、回答者によって多義的な解釈が可能になってしまった変数であろうと思われる。この表から因子負荷量が 0.5 以上（この値は恣意的に決めた）のものだけ残して変数ごとの質問内容も付記すると、下表が得られる。

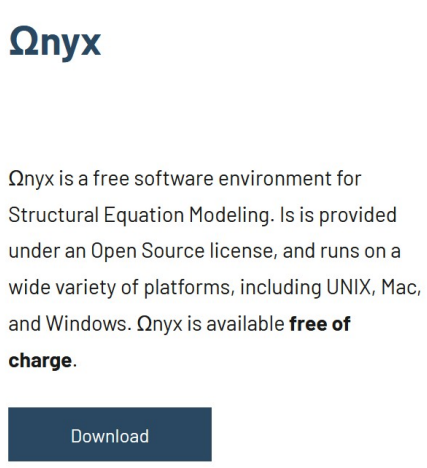
```
$$\vbox{\small
\begin{tabular}{lccccc}
\hline
変数              & MR2 & MR1 & MR3 & MR4 & MR5 \cr
\hline
Q01.紙リサイクル   &     &     &     &0.625&     \cr
Q03.容器リサイクル &     &     &     &0.670&     \cr
Q05.冷暖房控える   &     &     &     &     &0.516\cr
Q06.食材適量購入   &     &     &     &     &0.625\cr
Q11.太陽熱温水器   &     &     &0.581&     &     \cr
Q13.米のとぎ汁利用 &     &     &0.527&     &     \cr
Q15.塩ビラップ不買 &0.502&     &     &     &     \cr
Q16.石けん使う     &     &     &0.554&     &     \cr
Q18.除草殺虫剤不使用&0.687&     &     &     &     \cr
Q20.強力洗浄剤不使用&0.632&     &     &     &     \cr
Q21.有機溶剤不使用 &0.650&     &     &     &     \cr
Q22.有機農産物選好 &     &0.789&     &     &     \cr
Q23.地場農産物選好 &     &0.930&     &     &     \cr
\hline
```

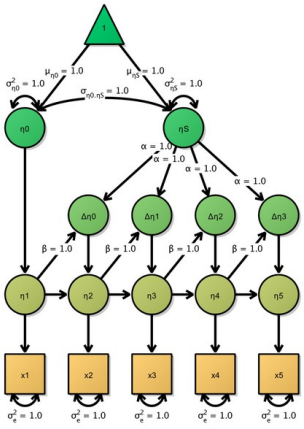
\end{tabular}}$$

もしこれを因子分析結果として採用し、下位尺度の得点の計算に使うならば、これらの変数だけを使って因子分析をやり直す必要があるが、本稿ではそこまで深入りしない。

Structural Equation Modeling (SEM)

In structural equation modeling (SEM), the relationships between multiple observed variables and unobserved latent factors (constructs), including the direction of influence, are modeled and then fitted to data. There are various packages for structural equation modeling in R, but **`sem`** and **`lavaan`** are well-known. For visualizing the results, the **`semPaths()`** function from the **`semPlot`** package is convenient. Since there are countless possibilities for relationships between variables, directly coding them can be challenging until you get used to it. Therefore, auxiliary software that generates analysis code automatically by drawing relationship diagrams using a GUI (Graphical User Interface) can be helpful[8]. There is a very famous supporting software for sem, Ωnyx[9], which is freely available.





Ωnyx is written in Java, so that installation of Java runtime is necessary. Oracle Open JDK is linked from the official website of Ωnyx[10], but I recommend rather Adoptium JDK[11].
If the Java runtime environment or development kit is installed, downloaded file with extension .jar such as onyx-1.0-1043.jar is executable by double-click. The user guide[12] is also available.

The basics of SEM

There are 5 rules, in principle, to draw the path diagram to show the relationships among observed variables and latent factors (constructs) shown below.

- Observed variables in rectangular box
- Latent factors in ellipse
- Draw arrows from the source of effects to the targets
- 2 exogenous variables with covariate relations to be connected by line with arrows at both edge
- Residual (error) variable in circle or only value

8 It might be efficient to generate the initial code from a GUI layout and then manually adjust it later.
9 https://onyx-sem.com/
10 https://openjdk.org/index.html
11 https://adoptium.net/temurin/releases/
12 https://onyx-sem.com/wp-content/uploads/2021/08/userguide.pdf

The fundamental approach to structural equation modeling is to establish models for each endogenous variable that is influenced within a path diagram. A model comprises both measurement equations and structural equations.
A measurement equation describes how a construct influences multiple observed variables (it can also be thought of as an equation expressing how a construct is measured by observed variables, hence the term "measurement equation"). It takes the form: Observed Variable 1 = Coefficient 1 × Construct 1 + Coefficient 2 × Construct 2 + ...
A structural equation expresses the influence relationships between variables. This can involve constructs influencing other constructs, observed variables influencing other observed variables, or observed variables influencing constructs.
Besides these, there are "covariate relationships" where errors (unique variances) of observed variables, not explained by other variables' influences, are thought to be related. The model will express these three types of relationships.
You can't just draw a path diagram out of thin air without any analysis. Typically, you'd hypothesize and model potential path diagrams by examining scatter plots, correlation matrices, performing exploratory factor analysis, or reviewing prior research. The syntax for drawing models varies depending on the package that implements structural equation modeling. In this text, we'll explain the lavaan and sem packages.The e-learning course in UCLA to learn SEM using lavaan is very informative[13].

Example of CFA to apply ecopoint data

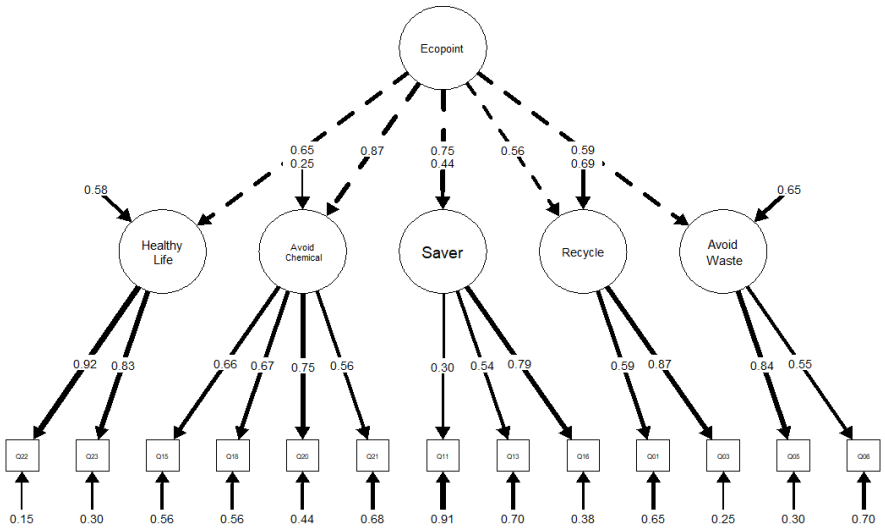
To get the converged solution, the result from EFA was slightly modified.

```
eco <- read.delim("ecopx.txt")
ecodata <- eco[, c(1, 3, 5, 6, 11, 13, 15, 16, 18, 20:23)+3]
C1 <- cor(ecodata)
library(sem)
M1 <- specifyEquations(text="
 Q22 = a1*HealthyLife
 Q23 = a2*HealthyLife
 Q18 = b1*AvoidChem
 Q20 = b2*AvoidChem
 Q21 = b3*AvoidChem
 Q15 = b4*AvoidChem
 Q11 = c1*Saver
 Q13 = c2*Saver
 Q16 = c3*Saver
 Q01 = d1*Recycle
 Q03 = d2*Recycle
 Q05 = e1*AvoidWaste
 Q06 = e2*AvoidWaste
 HealthyLife = 1*Ecopt
 AvoidChem = 1*Ecopt
 Saver = 1*Ecopt
 Recycle = 1*Ecopt
```

13 https://stats.oarc.ucla.edu/r/seminars/rsem/

```
 AvoidWaste = 1*Ecopt
 V(Ecopt) = 1
")
S1 <- sem(M1, C1, N=length(ecodata[, 1]))
print(S1)
summary(S1, fit.indices=c("GFI","AGFI","CFI","RMSEA"))
library(semPlot)
LBL <- c("Q22","Q23","Q15","Q18","Q20","Q21","Q11","Q13","Q16",
         "Q01","Q03","Q05","Q06",
         "Healthy\n Life","Avoid\n Chemical","Saver","Recycle",
         "Avoid\n Waste","Ecopoint")
semPaths(S1, what="stand", layout="tree", style="lisrel",
         shapeMan="rectangle", shapeLat="ellipse",
         sizeMan=3, residScale=9, posCol="black",
         negCol="red", fade=FALSE, edge.label.cex=0.8,
         nodeLabels=LBL)
```

By running this code, the path diagram below is obtained and the coefficients and goodness of fit indices are obtained.



The options in **`semPaths()`** are explained below (copied from help).

object A "semPlotModel" object or any of the input types that can be used in semPlotModel directly.
what What should the edges indicate in the path diagram? This function uses grepl to allow fuzzy matching and is not case sensitive. E.g., par will also match Parameters.
path, diagram or mod This will display the model as an unweighted network (gray edges by default).
est or par This will display the parameter estimates as weighted edges.
stand or std This will display the standardized parameter estimates, if available, as weighted edges.
eq or cons This is the same graph as path. except that parameters with equality constraints are now colored. Parameters with the same color are constrained to be equal.

col This will create an unweighted graph of the path diagram, where edges are colored with a mix of the colors of connected nodes.
whatLabels What should the edge labels indicate in the path diagram? This function uses grepl to allow fuzzy matching and is not case sensitive. E.g., par will also match Parameters. Default depends on the what argument, defaulting to the respective elements in the list below for values of what in the list above.
name, label, path or diagram This will display the edge names as labels.
est or par This will display the parameter estimate in edge labels.

stand or std
This will display the standardized parameter estimate in edge labels.

eq or cons
This will display the parameter number in edge labels. 0 indicates the parameter is fixed, parameters with the same parameter number are constrained to be equal.

no, omit, hide or invisible
Hides edge labels.

style
The style to use. Currently only indicates what the (residual) variances look like. Use "ram", "mx" or "OpenMx" for double headed selfloops and "lisrel" for single headed edges with no node as origin. Defaults to "ram" unless the input is a lisrel model.

layout
A string indicating how the nodes should be placed. Similar to the 'layout' argument in qgraph. Can be one of the following strings.

tree
The integrated tree-like layout. Places exogenous variables at the top and endogenous variables at the bottom. See 'details' for more details.

circle
The same layout as "tree", except that afterwards the horizontal levels of the layout are placed in circles. Especially useful for models with a large number of manifest variables and a relatively small number of latent variables.

spring
Calls the "spring" layout in qgraph, which uses the Fruchterman-reingold algorithm (Fruchterman & Reingold, 1991).

tree2
Calls the layout.reingold.tilford function from the igraph package (Csardi & Nepusz, 2006), which uses the Reingold-Tilford algorithm (Reingold & Tilford, 1981). Before calling the algorithm roots are chosen and a slightly modified version of the graph is used to produce consistent results. See 'details'.

circle2

The same layout as "tree2", except that afterwards the horizontal levels of the layout are placed in circles.

Other options
If the assigned value is not in this list it is sent to qgraph. This allows for manual specification of the layout as well as using functions found in the 'igraph; library.

intercepts
Logical, should intercepts be included in the path diagram?

residuals
Logical, should residuals (and variances) be included in the path diagram?

thresholds
Logical, should thresholds be included in the path diagram?

intStyle
Style of the intercepts. "multi" plots a separate unit vector node for each intercept and "single" plots a single unit vector node. Currently, "single" is not well supported and might lead to unexpected results.

rotation
An integer indicating the rotation of the layout when "tree" or "tree2" layout is used. 1, 2, 3 and 4 indicate that exogenous variables are placed at the top, left side, bottom and right side respectively.

curve
The curvature of the edges. In tree layouts this argument only curves the edges that are between nodes on the same level. e.g., correlations between exogenous manifest variables.

curvature
Sets the strength of scaling in curvature for curved edges at the same horizontal level in tree layouts. The curve will be set to curve + curvature * n / max(n), where n is the number of nodes in between the two connected nodes.

nCharNodes
Number of characters to abbreviate node labels to (using abbreviate). Set to 0 to omit abbreviation.

nCharEdges
Number of characters to abbreviate edge labels to (using abbreviate). Set to 0 to omit abbreviation.

sizeMan
Width of the manifest nodes, sent to the 'vsize' argument in qgraph.

sizeLat
Width of the latent nodes, sent to the 'vsize' argument in qgraph.

sizeInt
Width of the unit vector nodes, sent to the 'vsize' argument in qgraph.

sizeMan2

Height of the manifest nodes, sent to the 'vsize2' argument in qgraph.

sizeLat2
Height of the latent nodes, sent to the 'vsize2' argument in qgraph.

sizeInt2
Height of the unit vector nodes, sent to the 'vsize2' argument in qgraph.

shapeMan
Shape of the manifest nodes, sent to the 'shape' argument in qgraph. Defaults to "square" or "rectangle" if width and height differ.

shapeLat
Shape of the latent nodes, sent to the 'shape' argument in qgraph. Defaults to "circle" or "ellipse" if width and height differ.

shapeInt
Shape of the constant nodes, sent to the 'shape' argument in qgraph. Defaults to "triangle".

ask
Specifies the 'ask' parameter in par. Defaults to TRUE if multiple groups are in the model.

mar
Same as the 'mar' argument in qgraph. By default this argument is based on the values of 'rotation', 'style' and 'title'.

title
Logical, should titles be plotted of the group names above each plot?

title.color
Color of the titles.

title.adj
Adjustment of title as used by 'adj' in par.

title.line
Line of title as used by 'line' in title.

title.cex
Size of title as used by 'cex.main' in par.

include
Integer vector indicating which groups should be included in the output. e.g., to only plot a diagram for the first group use include = 1.

combineGroups
Logical. If TRUE all groups are combined in the same path diagram.

manifests

A character vector in which every element is the name of a manifest variable in the model. This argument can be used to overwrite the order in which nodes are plotted in the graph if reorder = FALSE

latents
A character vector in which every element is the name of a latent variable in the model. This argument can be used to overwrite the order in which nodes are plotted in the graph if reorder = FALSE

groups
Groups nodes that should be colored the same, similar to the 'groups' argument in qgraph with a few exceptions. Should be a list containing in each element the names (instead of numbers as in qgraph) of nodes that belong together. Nodes that are indicated to belong to a group will be assigned the same color, as given by the 'color' argument. Nodes not belonging to a group will be assigned the color "", which indicates that they will inherit a mix of the colors of connected nodes (or white, if no connected nodes are colored.)

In addition, this argument can be assigned a single character: "manifests", "latents" or "both" to make a single group for each manifest, latent or both manifest and latent variables. e.g., groups = "latents" will color each latent variable uniquely, and color all manifest variables a mixture of the colors of latents they load on.

color
Controls the color of nodes. Similar to 'color' in qgraph. A color vector indicating the color for each group, a single color character indicating the color for all nodes or a color vector indicating the color for each node separately. Can also be a list contaning one or more of the following elements (using fuzzy matching):

man
The colors for manifest nodes

lat
The colors for latent nodes

int
The color for intercepts

residScale
The size of residual edges if style = "lisrel". Defaults to two times the value of 'sizeMan'.

gui
Not yet implemented.

allVars
Logical. If TRUE all variables are plotted in the path diagrams for each group. If FALSE only variables are plotted that are used in the group.

edge.color
A value indicating the color of all edges or a vector indicating the color of each edge. Useful for manually overwriting edge colors.

reorder
Logical. Should manifest variables be reordered to be near latent factors they are most connected to in the "tree" layout? If FALSE manifest variables are placed in the order they appear in the Pars.

structural
Logical. Set this to TRUE to only show the structural model (omit all manifest variables.)

ThreshAtSide
Logical. If TRUE, thresholds are plotted as small lines at the side of manifest nodes, otherwise they are plotted as lines inside the nodes.

thresholdColor
Color of the threshold lines. Defaults to "black"

thresholdSize
Size of threshold bars relative to the size of the node.

fixedStyle
A vector of length one or two specifying the color and line type (same as 'lty' in par) of fixed parameters. Can be both character and numeric. If one of the elements encodes a color it is used to overwrite the color of fixed edges, and if an element can be coerced to a numeric it is used to encode the line type.

For example, fixedStyle = c("red",3) specifies that all fixed parameters should be visualized with a red edge with lty=3

freeStyle
Same as 'fixedStyle' but for free parameters instead.

as.expression
A character vector indicating which labels should be treated as an expression, so that mathematical notation and Greek letters can be used in the path diagram. If this vector contains "nodes" all node labels are converted to expressions, and if this vector contains "edges" all node labels are converted to expressions. Defaults to "edges" only if the input is a Lisrel model.

optimizeLatRes
Logical. If this is TRUE, the angle of the incoming residuals on latent variables is attempted to be optimally chosen so its position conflicts with the least amount of connected edges.

inheritColor
Logical, should uncolored nodes obtain a mix of connected colored nodes? Defaults to TRUE.

levels
A numeric vector usually of length 4. Controls the relative vertical position of variable levels (exogenous and endogenous latents and manifests) under default rotation in tree and circle layouts. This can be used to control the spacing between these levels. e.g., c(1,5,6,7) will create more space between endogenous manifests and latents.

nodeLabels

A vector or list to manually overwrite the node labels. Can include expressions.

edgeLabels
A vector or list to manually overwrite the edge labels. Can include expressions.

pastel
Logical, should default colors (for groups or edge equality constraints) be chosen from pastel colors? If TRUE then rainbow_hcl is used.

rainbowStart
A number between 0 and 1 indicating the offset used in rainbow functions for default node coloring.

exoVar
Should variances of truely exogenous variables (no incomming directed edge) be plotted? Defaults to TRUE unless style = "lisrel".

intAtSide
Logical to control if intercepts should be plotted to the side of manifest nodes or at the bottom/top. Defaults only to FALSE if 'residuals=FALSE'.

springLevels
Logical indicating if the placement on horizontal levels with tree3 layout should be determined by a force embedded algorithm.

nDigits
Number of digits to round numeric values to.

exoCov
Should covariances between truely exogenous variables (no incomming directed edge) be plotted? Defaults to TRUE.

centerLevels
Only used if layout is set to "tree2", should each level be centered? Defaults to TRUE

panelGroups
Logical to automatically create a panel plot of multiple group models. Defaults to FALSE.

layoutSplit
Logical that can be used to split computing of layout between structural and measurment models. This is very useful in more complicated models where the structural part is best shown by using a spring layout.

measurementLayout
Logical indicating the layout algorithm to use for measurement models if layoutSplit = TRUE (the structural model will obtain a layout given by the layout argument).

subScale
Width of submodels (measurment models) if layoutSplit = TRUE.

subScale2
Height of submodels (measurment models) if layoutSplit = TRUE.

subRes
Integer indicating the resolution of which measurment models can be rotated around their corresponding latent variable. The default, 4, indicates that they can be placed only to polar coordinates. Set to 360 to allow every angle of rotation.

subLinks
Vector of variables to link to. Currently not well supported so avoid using this argument.

modelOpts
A lists containing arguments sent to semPlotModel in case the input is not of class semPlotModel.

curveAdjacent
What edges between adjacent horizontal nodes be curved? Can be '<->' or 'cov' to indicate bidirectional covariances, '->' or 'reg' for directed regressions or a vector containing both.

edge.label.cex
Controls the font size of the edge labels. Same as in qgraph except that the default is now 0.8.

cardinal
Should edges in a tree layout connect to the four cardinal points of one of the borders of the node rather than point to the center of the node? Can be set to TRUE or "all" to enamble this behavior for all edges and FALSE or "none" to disable this behavior for all edges. Alternatively a vector with strings can be specified in which each string specifies a certain group of edges. Fuzzy matching is used on the strings "exo" for edges with the first node being exogenous (or indicator of exogenous latent), endo for edges with first node being endogeonous, manifest for edges connected to any manifest node, latent for edges connected to any latent node, cov for covariances, reg for regressions, load for factor-loadings, source for only the start of an edge and end for only the end of a node. These strings can be combined at will. For example, cardinal = c("exo cov","load end") (the default) or equivelantly cardinal = c("exogenous covariances","source of loadings") will only cardenalize the edges that represent exogenous covariances or the end of factor loadings.

equalizeManifests
Logical. Should the distances between manifest nodes in the tree1 layout be equalized? Defaults to TRUE

covAtResiduals
Logical, should covariances be drawn at the start of residuals when style="lisrel" is used? Defaults to TRUE.

bifactor
A string vector containing the name(s) of the general bifactor(s). This will automatically create a bifactor plot.

optimPoints
A vector of radians residuals can optimize to if optimizeLatRes = TRUE

...
Arguments sent to the qgraph function. These arguments can further control the output of the graph. Some usefull arguments in drawing path diagrams are:

edge.width
Scales the edge width and arrow size of the plot. These can also be manually set using 'esize' and 'asize'.

node.width
Scales the width of nodes and also the height if shapes circle and square are used. Can also be a vector with scalar for each node.

node.height
Scales the height of nodes. Can also be a vector with scalar for each node. Not used with circle and square shapes.

esize
Size of the largest edge (or what it would be if there was an edge with weight maximum). Defaults to: max((-1/72)*(nNodes)+5.35,1) for weighted graphs and 2 for unweighted graphs. In directed graphs these values are halved.

asize
Size of the arrowhead. Defaults to 2 for graphs with more than 10 nodes and 2 to smaller graphs.

minimum
Edges with absolute weights under this value are omitted. Defaults to 0 for graphs with less than 50 nodes or 0.1 for larger graphs.

maximum
qgraph regards the highest of the maximum or highest absolute edge weight as the highest weight to scale the edge widths too. To compare several graphs, set this argument to a higher value than any edge weight in the graphs (typically 1 for correlations).

cut
In weighted graphs, this argument can be used to cut the scaling of edges in width and color saturation. Edges with absolute weights over this value will have the strongest color intensity and become wider the stronger they are, and edges with absolute weights under this value will have the smallest width and become vaguer the weaker the weight. If this is set to NULL, no cutoff is used and all edges vary in width and color. Defaults to NULL for graphs with less then 50 nodes and 0.3 to larger graphs.

details
Logical indicating if minimum, maximum and cutoff score should be printed under the graph. Defaults to FALSE.

mar
A vector of the form c(bottom, left, top, right) which gives the margins. Works similar to the argument in par(). Defaults to c(3,3,3,3)

filetype

A character containing the file type to save the output in. "R" outputs in a new R window, "pdf" creates a pdf file. "svg" creates a svg file (requires RSVGTipsDevice). "tex" creates LaTeX code for the graph (requires tikzDevice). 'jpg', 'tiff' and 'png' can also be used. If this is given any other string (e.g. filetype="") no device is opened. Defaults to 'R' if the current device is the NULL-device or no new device if there already is an open device. A function such as x11 can also be used

filename
Name of the file without extension

width
Width of the plot, in inches

height
Height of the plot, in inches

normalize
Logical, should the plot be normalized to the plot size. If TRUE (default) border width, vertex size, edge width and arrow sizes are adjusted to look the same for all sizes of the plot, corresponding to what they would look in a 7 by 7 inches plot if normalize is FALSE.

DoNotPlot
Runs qgraph but does not plot. Useful for saving the output (i.e. layout) without plotting

plot
Logical. Should a new plot be made? Defaults to TRUE. Set to FALSE to add the graph to the existing plot.

rescale
Logical. Defines if the layout should be rescaled to fit the -1 to 1 x and y area. Defaults to TRUE. Can best be used in combination with plot=FALSE.

label.cex
Scalar on the label size.

label.color
Character containing the color of the labels, defaults to "black"

borders
Logical indicating if borders should be plotted, defaults to TRUE.

border.color
Color vector indicating colors of the borders. Is repeated if length is equal to 1. Defaults to "black"

border.width
Controls the width of the border. Defaults to 2 and is comparable to 'lwd' argument in 'points'.

polygonList
A list contaning named lists for each element to include polygons to lookup in the shape arguments. Each element must be named as they are used in shape and contain a list with elements x and y

contaning the coordinates of the polygon. By default ellipse and heart are added to this list. These polygons are scaled according to vsize and vsize2

vTrans
Transparency of the nodes, must be an integer between 0 and 255, 255 indicating no transparency. Defaults to 255

label.prop
Controls the proportion of the width of the node that the label rescales to. Defaults to 0. 9.

label.norm
A single string that is used to normalize label size. If the width of the label is lower than the width of the hypothetical label given by this argument the width of label given by this argument is used instead. Defaults to "OOO" so that every label up to three characters has the same fontsize.

label.scale
Logical indicating if labels should be scaled to fit the node. Defaults to TRUE.

label.font
Integer specifying the label font of nodes. Can be a vector with value for each node

posCol
Color of positive edges. Can be a vector of two to indicate color of edges under 'cut' value and color of edges over 'cut' value. If 'fade is set to TRUE the first color will be faded the weaker the edge weight is. If this is only one element this color will also be used for edges stronger than the 'cut' value. Defaults to c("#009900","darkgreen")

negCol
Color of negative edges. Can be a vector of two to indicate color of edges under 'cut' value and color of edges over 'cut' value. If 'fade is set to TRUE the first color will be faded the weaker the edge weight is. If this is only one element this color will also be used for edges stronger than the 'cut' value. Defaults to c("#BF0000","red")

unCol
Color to indicate the default edge color of unweighted graphs. Defaults to "#808080".

colFactor
Exponent of transformation in color intensity of relative strength. Defaults to 1 for linear behavior.

trans
In weighted graphs: logical indicating if the edges should fade to white (FALSE) or become more transparent (TRUE; use this only if you use a background). In directed graphs this is a value between 0 and 1 indicating the level of transparency. (also used as 'transparency')

fade
if TRUE (default) and if 'edge.color' is assigned, transparency will be added to edges that are not transparent (or for which no transparency has been assigned) relative to the edge strength, similar if 'trans' is set to TRUE.

loop
This can be used to scale the size of the loop. defaults to 1.

curvePivot
Quantile to pivot curves on. This can be used to, rather than round edges, make straight edges as curves with "knicks" in them. Can be logical or numeric. FALSE (default) indicates no pivoting in the curved edges, a number indicates the quantile (and one minus this value as quantile) on which to pivot curved edges and TRUE indicates a value of 0.1.

curvePivotShape
The shape of the curve around the pivots, as used in xspline. Defaults to 0.25.

edge.label.bg
Either a logical or character vector/matrix. Indicates the background behind edge labels. If TRUE (default) a white background is plotted behind each edge label. If FALSE no background is plotted behind edge labels. Can also be a single color character, a vector or matrix of color vectors for each edge.

edge.label.position
Vetor of numbers between 0 and 1 controlling the relative position of each edge label. Defaults to 0.5 for placing edge labels at the middle of the edge.

edge.label.font
Integer specifying the label font of edges. Can be a vector or matrix with value for each node

layout.par
A list of arguments passed to qgraph.layout.fruchtermanreingold when layout="spring" or to an igraph function when such a function is assigned to 'layout'

bg
If this is TRUE, a background is plotted in which node colors cast a light of that color on a black background. Can also be a character containing the color of the background Defaults to FALSE

bgcontrol
The higher this is, the less light each node gives if bg=TRUE. Defaults to 6.

bgres
square root of the number of pixels used in bg=TRUE, defaults to 100.

pty
See 'par'

font
Integer specifying the default font for node and edge labels

arrows
A logical indicating if arrows should be drawn, or a number indicating how much arrows should be drawn on each edge. If this is TRUE, a simple arrow is plotted, if this is a number, arrows are put in the middle of the edges.

arrowAngle
Angle of the arrowhead, in radians. Defaults to pi/8 for unweighted graphs and pi/4 for weighted graphs.

asize
Size of the arrowhead. Defaults to 2 for graphs with more than 10 nodes and 2 to smaller graphs.

open
Logical indicating if open (TRUE) or closed (FALSE) arrowheads should be drawn.

weighted
Logical that can be used to force either a weighted graph (TRUE) or an unweighted graph(FALSE).

XKCD
If set to TRUE the graph is plotted in XKCD style based on
http://stackoverflow.com/a/12680841/567015.

\item[what] 矢印の上に何を表示するかを指定する。\verb!"stand"!だと標準化したパラメータ推定値が表示される。標準化されていないパラメータ推定値を表示したい場合は\verb!"est"!とする。デフォルトではパラメータ名が表示される。
\item[layout] 関連図の配置パタンを指定する。デフォルトは\verb!"tree"!だが、\verb!"spring"!とすると下図のような不規則な配置になる。円環状に配置したいときは\verb!"circle"!にする。
\item[style] 誤差分散の表示スタイルを指定する。デフォルトでは枠付きの円状の両向き矢印だが、\verb!"lisrel"!と指定すると、枠無しで変数に向かう矢印が表示される
\item[shapeMan] 観測変数の枠のスタイルで正方形か長方形か選べると書かれているが、\verb!"rectangle"!と指定しても下図のように正方形になってしまった。
\item[shapeLat] 潜在因子の枠のスタイルで正円か楕円か選べると書かれているが、\verb!"ellipse"!と指定しても下図のように正円になってしまった。
\item[sizeMan] 観測変数の枠サイズ
\item[residScale] 残差の表示サイズ（デフォルトは観測変数の枠サイズの 2 倍）
\item[posCol] パラメータ推定値が正な矢印の色。デフォルトは緑。
\item[negCol] パラメータ推定値が負な矢印の色。デフォルトは赤。
\item[fade] デフォルトでは\verb!TRUE!になっていて、絶対値がゼロに近いパラメータや矢印ほど薄い色で表示される（透過性が高くなる）。すべての関連を同じ濃さで表示したいときは\verb!FALSE!にする。
\item[edge.label.cex] パラメータの文字サイズを基準フォントサイズの何倍にするか。デフォルトは 0.8 倍。
\item[nodeLabels] 観測変数名と潜在因子名を文字列ベクトルとして与える。このオプションを指定しないと、モデルに与えた変数名が短縮されて表示される。

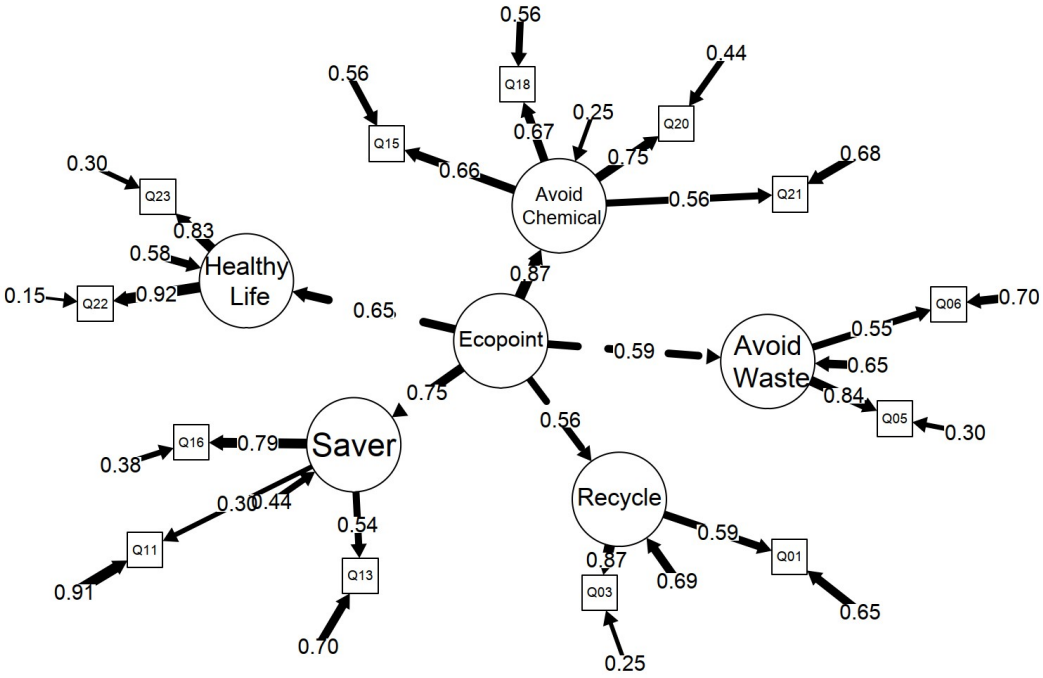
AGFI (adjusted goodness of fit index) is less than 0.9 and RMSEA is large, so that this model doesn’t explain the factor structure enough.

```
Model Chisquare =  205.5724   Df =  60 Pr(>Chisq) = 7.976845e-18
Goodness-of-fit index =  0.8969044
Adjusted goodness-of-fit index =  0.8436383
RMSEA index =  0.08889854   90% CI: (0.0757616, 0.102366)
```

```
 Bentler CFI =  0.880862
 Normalized Residuals
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-5.372000 -0.567000 -0.000001 -0.137700  0.536800  4.030000
 R-square for Endogenous Variables
HealthyLife         Q22         Q23   AvoidChem         Q18
     0.4201      0.8547      0.6972      0.7492      0.4430
        Q20         Q21         Q15       Saver         Q11
     0.5642      0.3155      0.4407      0.5626      0.0918
        Q13         Q16     Recycle         Q01         Q03
     0.2966      0.6205      0.3132      0.3486      0.7521
 AvoidWaste         Q05         Q06
     0.3509      0.7026      0.2977

 Parameter Estimates
                Estimate  Std Error  z value   Pr(>|z|)
a1              0.5992096 0.06007680  9.974060 1.979679e-23
a2              0.5411689 0.06140249  8.813469 1.213322e-18
b1              0.5760801 0.05832424  9.877198 5.227435e-23
b2              0.6501434 0.05884422 11.048553 2.227771e-28
b3              0.4861677 0.05737281  8.473835 2.374427e-17
b4              0.5746401 0.05830885  9.855110 6.514555e-23
c1              0.2272601 0.05324816  4.267942 1.972844e-05
c2              0.4084667 0.06124531  6.669354 2.569313e-11
c3              0.5908261 0.06013406  9.825149 8.774419e-23
d1              0.3304310 0.06406010  5.158141 2.494135e-07
d2              0.4853548 0.06162453  7.876000 3.380271e-15
e1              0.4965253 0.06148035  8.076162 6.683681e-16
e2              0.3231900 0.06422009  5.032537 4.840307e-07
V[HealthyLife] 1.3805665 0.36574314  3.774689 1.602075e-04
V[Q22]          0.1452525 0.05988679  2.425452 1.528935e-02
V[Q23]          0.3028184 0.05377719  5.630982 1.791865e-08
V[AvoidChem]    0.3347199 0.13876381  2.412156 1.585850e-02
V[Q18]          0.5570489 0.05555878 10.026299 1.168138e-23
V[Q20]          0.4358320 0.05116138  8.518768 1.612593e-17
V[Q21]          0.6845269 0.06201481 11.038118 2.502185e-28
V[Q15]          0.5592605 0.05565876 10.048023 9.372906e-24
V[Saver]        0.7774949 0.28505985  2.727480 6.382022e-03
V[Q11]          0.9081974 0.07610313 11.933771 7.891617e-33
V[Q13]          0.7034338 0.06866133 10.244978 1.246531e-24
V[Q16]          0.3795202 0.08214514  4.620117 3.835235e-06
V[Recycle]      2.1928126 0.75648576  2.898683 3.747332e-03
V[Q01]          0.6513939 0.07657334  8.506797 1.788036e-17
V[Q03]          0.2478714 0.12176585  2.035640 4.178652e-02
V[AvoidWaste]   1.8499837 0.66311145  2.789853 5.273192e-03
V[Q05]          0.2973725 0.12183788  2.440723 1.465789e-02
V[Q06]          0.7023140 0.07599017  9.242170 2.415488e-20

a1              Q22 <--- HealthyLife
```

```
a2              Q23 <--- HealthyLife
b1              Q18 <--- AvoidChem
b2              Q20 <--- AvoidChem
b3              Q21 <--- AvoidChem
b4              Q15 <--- AvoidChem
c1              Q11 <--- Saver
c2              Q13 <--- Saver
c3              Q16 <--- Saver
d1              Q01 <--- Recycle
d2              Q03 <--- Recycle
e1              Q05 <--- AvoidWaste
e2              Q06 <--- AvoidWaste
V[HealthyLife] HealthyLife <--> HealthyLife
V[Q22]          Q22 <--> Q22
V[Q23]          Q23 <--> Q23
V[AvoidChem]   AvoidChem <--> AvoidChem
V[Q18]          Q18 <--> Q18
V[Q20]          Q20 <--> Q20
V[Q21]          Q21 <--> Q21
V[Q15]          Q15 <--> Q15
V[Saver]        Saver <--> Saver
V[Q11]          Q11 <--> Q11
V[Q13]          Q13 <--> Q13
V[Q16]          Q16 <--> Q16
V[Recycle]     Recycle <--> Recycle
V[Q01]          Q01 <--> Q01
V[Q03]          Q03 <--> Q03
V[AvoidWaste]  AvoidWaste <--> AvoidWaste
V[Q05]          Q05 <--> Q05
V[Q06]          Q06 <--> Q06
```

When the ”spring” instead of “tree” is given to the option layout, the graph becomes below.



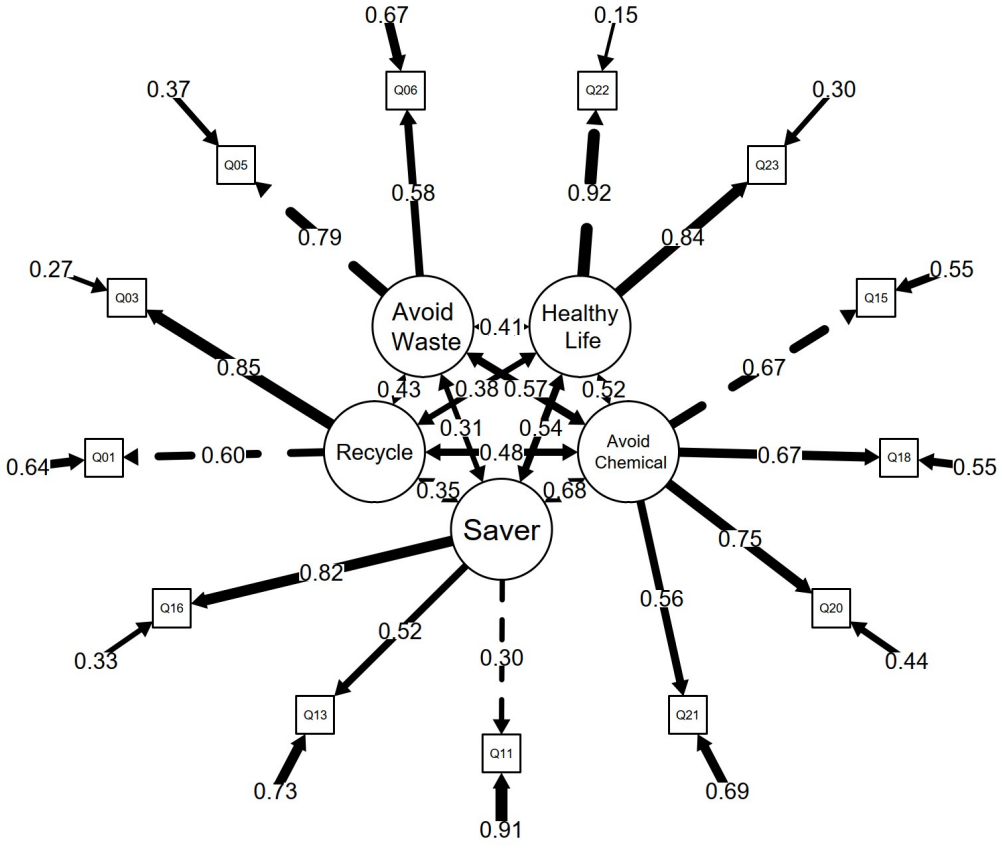
Using lavaan package

The ways of model specification in **lavaan** and **sem** are different. The **lavaan** package uses =~ for regression equation, ~ for structural equation, and ~~ for covariate relationship. In addition, the **lavaan** package implicitly assume many coefficients and paths unless strictly specified, and thus the model specification becomes shorter than **sem** package. If the same factor structure is written in **lavaan**, the code[14] follows.

```
ecodata <- eco[, c(1, 3, 5, 6, 11, 13, 15, 16, 18, 20:23)+3]
ecodata <- subset(ecodata, complete.cases(ecodata))
library(lavaan)
M1 <- 'HealthyLife =~ Q22 + Q23
AvoidChem =~ Q18 + Q20 + Q21 + Q15
Saver =~ Q11 + Q13 + Q16
Recycle =~ Q01 + Q03
AvoidWaste =~ Q05 + Q06'
S1 <- sem(model=M1, data=ecodata, estimator="ML")
summary(object=S1, fit.measure=TRUE)
library(semTools)
reliability(S1)
```

14 https://minato.sip21c.org/advanced-statistics/ecolavaan.R

```
library(semPlot)
LBL <- c("Q22","Q23","Q15","Q18","Q20","Q21","Q11","Q13","Q16",
         "Q01","Q03","Q05","Q06","Healthy\n Life","Avoid\n Chemical",
         "Saver","Recycle","Avoid\n Waste")
semPaths(S1, what="stand", layout="circle", style="lisrel",
       shapeMan="rectangle",
       shapeLat="ellipse",sizeMan=3, residScale=9, posCol="black",
       negCol="red", fade=FALSE, edge.label.cex=0.8,nodeLabels=LBL)
```



As shown below, CFI was slightly smaller than 0.9 and RMSEA was 0.090, which was statistically significantly larger than 0.05, so that the goodness of fit was not enough.

```
lavaan (0.5-20) converged normally after 73 iterations
  Number of observations                           308
  Estimator                                         ML
  Minimum Function Test Statistic              191.644
```

```
  Degrees of freedom                                55
  P-value (Chi-square)                           0.000

Model test baseline model:
  Minimum Function Test Statistic             1304.114
  Degrees of freedom                                78
  P-value                                        0.000

User model versus baseline model:
  Comparative Fit Index (CFI)                    0.889
  Tucker-Lewis Index (TLI)                       0.842

Loglikelihood and Information Criteria:
  Loglikelihood user model (H0)              -5904.159
  Loglikelihood unrestricted model (H1)      -5808.337

  Number of free parameters                         36
  Akaike (AIC)                               11880.318
  Bayesian (BIC)                             12014.601
  Sample-size adjusted Bayesian (BIC)        11900.424

Root Mean Square Error of Approximation:
  RMSEA                                          0.090
  90 Percent Confidence Interval          0.076  0.104
  P-value RMSEA <= 0.05                          0.000

Standardized Root Mean Square Residual:
  SRMR                                           0.071

Parameter Estimates:
  Information                                 Expected
  Standard Errors                             Standard

Latent Variables:
                   Estimate  Std.Err  Z-value  P(>|z|)
  HealthyLife =~
    Q22               1.000
    Q23               0.608    0.048   12.743    0.000
  AvoidChem =~
    Q15               1.000
    Q18               0.610    0.064    9.543    0.000
    Q20               0.600    0.058   10.304    0.000
    Q21               0.348    0.042    8.260    0.000
  Saver =~
    Q11               1.000
    Q13               0.777    0.186    4.187    0.000
    Q16               2.560    0.589    4.343    0.000
  Recycle =~
    Q01               1.000
```

```
    Q03               0.838    0.140    5.972    0.000
  AvoidWaste =~
    Q05               1.000
    Q06               0.654    0.107    6.124    0.000

Covariances:
                   Estimate  Std.Err  Z-value  P(>|z|)
  HealthyLife ~~
    AvoidChem         0.751    0.120    6.255    0.000
    Saver             0.201    0.053    3.823    0.000
    Recycle           0.498    0.117    4.256    0.000
    AvoidWaste        0.294    0.057    5.203    0.000
  AvoidChem ~~
    Saver             0.472    0.123    3.853    0.000
    Recycle           1.167    0.262    4.455    0.000
    AvoidWaste        0.761    0.127    5.970    0.000
  Saver ~~
    Recycle           0.221    0.076    2.923    0.003
    AvoidWaste        0.107    0.037    2.879    0.004
  Recycle ~~
    AvoidWaste        0.515    0.124    4.163    0.000

Variances:
                   Estimate  Std.Err  Z-value  P(>|z|)
    Q22               0.133    0.053    2.520    0.012
    Q23               0.124    0.022    5.752    0.000
    Q15               3.345    0.333   10.049    0.000
    Q18               1.250    0.124   10.063    0.000
    Q20               0.774    0.089    8.717    0.000
    Q21               0.720    0.065   11.096    0.000
    Q11               1.870    0.155   12.026    0.000
    Q13               0.300    0.028   10.759    0.000
    Q16               0.576    0.148    3.886    0.000
    Q01               3.858    0.454    8.497    0.000
    Q03               0.571    0.237    2.407    0.016
    Q05               0.398    0.104    3.823    0.000
    Q06               0.568    0.062    9.103    0.000
    HealthyLife       0.775    0.089    8.718    0.000
    AvoidChem         2.702    0.450    5.997    0.000
    Saver             0.180    0.078    2.317    0.021
    Recycle           2.172    0.499    4.353    0.000
    AvoidWaste        0.667    0.127    5.243    0.000
```

The reliability function in the semTools package provides Cronbach's alpha, McDonald's omega (a.k.a. Composite Reliability = CR), omega2 (unconditional reliability), omega3 (hierarchical omega) and AVE (Average Variance Extracted, suggesting convergent validity if AVE>0.5), as shown below. This result suggests HealthyLife and AvoidChem can be used as subscale of ecopoint, but all other subscales are not enough.

```
       HealthyLife AvoidChem     Saver   Recycle AvoidWaste     total
alpha    0.8338687 0.7006139 0.5190605 0.6188560  0.6249920 0.7751874
omega    0.8863942 0.7436756 0.5526472 0.6236869  0.6536104 0.8495406
omega2   0.8863942 0.7436756 0.5526472 0.6236869  0.6536104 0.8495406
omega3   0.8863945 0.7557620 0.5260580 0.6236864  0.6536105 0.8723029
avevar   0.8051577 0.4510670 0.3488440 0.4550706  0.4961635 0.4568644
```

The code https://minato.sip21c.org/advanced-statistics/ecolavaan2.R assumed the effect from total ecopoint on each subscale, but the results were almost same.

Using sem package, in reference to the text by Prof. John Fox

The **sem** package has been developed and maintained by Prof. John Fox, who wrote the tutorial text[15], and opened it to the public. The tutorial document explains not only SEM but also 2 stage least squares regression using **tsls()** function, where the regression model includes instrumental variables.
The tutorial includes the explanation for (1) structural equation model including potential exogenous variable (exogenous means the independent variable, free from residual variance in the model to explain the dependent variable) and endogenous variable (also independent variable in the model, but correlates with residual variance), (2) structural equation model where the observed variables are categorical variable, using the **hetcor()** function in the **polycor** package. To follow this tutorial, we need to installl **polycor** package too.

Typical SEM (1)

The code[16] given in the tutorial performs an analysis from a variance-covariance matrix, not from raw data. While variance-covariance matrices can be used as input for both the **sem** and **lavaan** packages, a correlation matrix does not seem to be intended as input for the **lavaan** package. The original source of this data is Wheaton et al. (1977)[17]. Professor John Fox has also published an exercise[18] using this data.
According to the original text by Wheaton et al. (1977), this data comes from a longitudinal study conducted over three periods (1966, 1967, and 1971) in areas where the Jones & Laughlin Steel Company was constructing a cold-rolling mill in rural Illinois, and in control areas where no such development was occurring, to investigate the impact of industrial development[19]. Survey results from 932 individuals across both regions were obtained at these three time points.
The variables Professor Fox used in this analysis are: Anomia scale scores from 1967 and 1971 (**Anomia67** and **Anomia71**)[20], Powerlessness scale scores from 1967 and 1971 (**Powerless67** and **Powerless71**)[21], and six variables in total: Education (**Education**, years of completed

15 https://socserv.socsci.mcmaster.ca/jfox/Misc/sem/SEM-paper.pdf
16 https://minato.sip21c.org/advanced-statistics/sem1.R
17 The original paper is available as a PDF file from https://www.statmodel.com/bmuthen/articles/Article_001.pdf
18 https://statmath.wu.ac.at/courses/StatsWithR/Exercises-5.pdf](https://statmath.wu.ac.at/courses/StatsWithR/Exercises-5.pdf
19 Details of the survey are referenced as Summers et al. (1969). A scanned version of the original report is available in full at [https://eric.ed.gov/?id=ED048953].
20 According to dictionaries, "anomia" refers to anomic aphasia, while "anomie" refers to anomie, valuelessness, or normlessness. However, the papers clearly deal with anomie as defined by Durkheim, and anomic aphasia is irrelevant.
21 Wheaton et al. (1977) states that powerlessness was developed by Summers in this study, but no further details are provided. With a Cronbach's alpha coefficient of 0.64, the reliability of the scale score is insufficient.

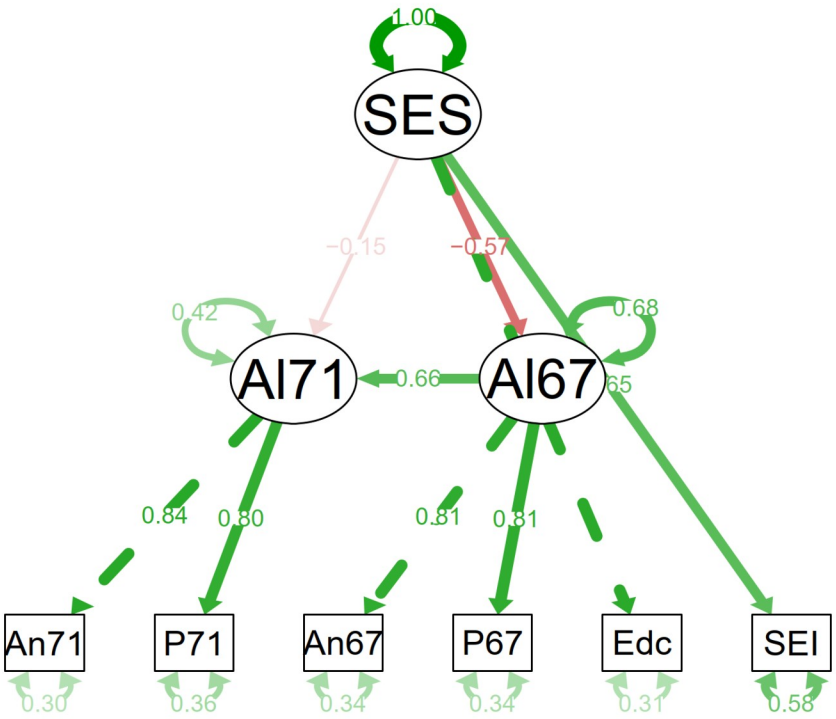
schooling) and Socioeconomic Index (**SEI**, known as Duncan's SEI, a score of occupational prestige based on occupational groups in the census, used as an index indicating the average socioeconomic status of that occupation).
It is stated that the Anomia and Powerlessness scale scores were originally subscales of an alienation scale. However, in Professor Fox's analysis, the alienation scale itself is not used as a numerical value; instead, it is included in the model as latent factors for each survey year (**Alienation67** and **Alienation71**). Furthermore, a latent factor named **SES** is hypothesized to underlie **Education** and **SEI** to represent socioeconomic status.
According to the original text by Wheaton et al. (1977) and Srole (1956) cited therein, the Anomia scale score is calculated from responses to five questions[22], where the result score was the number of "Agree".

```
library(sem)
mod.wh.1 <- specifyModel(text="
Alienation67 -> Anomia67, NA, 1
Alienation67 -> Powerless67, lam1, NA
Alienation71 -> Anomia71, NA, 1
Alienation71 -> Powerless71, lam2, NA
SES -> Education, NA, 1
SES -> SEI, lam3, NA
Alienation67 -> Alienation71, beta, NA
SES -> Alienation67, gam1, NA
SES -> Alienation71, gam2, NA
SES <-> SES, phi, NA
Alienation67 <-> Alienation67, psi1, NA
Alienation71 <-> Alienation71, psi2, NA
Anomia67 <-> Anomia67, the11, NA
Powerless67 <-> Powerless67, the22, NA
Anomia71 <-> Anomia71, the33, NA
Powerless71 <-> Powerless71, the44, NA
Education <-> Education, thd1, NA
SEI <-> SEI, thd2, NA
")
S.wh <- matrix(c(
11.834,0,0,0,0,0,
6.947,9.364,0,0,0,0,
6.819,5.091,12.532,0,0,0,
4.783,5.028,7.495,9.986,0,0,
-3.839,-3.889,-3.841,-3.625,9.610,0,
-21.899,-18.831,-21.748,-18.775,35.522,450.288),
6,6,byrow=TRUE)
rownames(S.wh) <- colnames(S.wh) <-
c('Anomia67','Powerless67','Anomia71','Powerless71','Education','SEI')

sem.wh.1 <- sem(mod.wh.1, S.wh, N=932)
summary(sem.wh.1, fit.indices=c("GFI","AGFI","CFI","RMSEA"))
library(semPlot)
semPaths(sem.wh.1, what="stand")
```

22 "There's little use of writing to public officials because often they aren't really interested in the problems of the average man.", "Nowadays a person has to live pretty much for today and let tomorrow take care of itself.", "In spite of what some people say, the lot of the average man is getting worse, not better.", "It's hardly fair to bring children into the world with the way things look for the future.", "These days a person doesn't really know whom he can count on."

The result follows. The resulted AGFI was 0.91, not so bad.

```
 Model Chisquare =  71.46973   Df =  6 Pr(>Chisq) = 2.041707e-13
 Goodness-of-fit index =  0.9751676
 Adjusted goodness-of-fit index =  0.9130866
 RMSEA index =  0.1082604   90% CI: (0.08658466, 0.1314454)
 Bentler CFI =  0.969066

 Normalized Residuals
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.2580000 -0.2118000 -0.0000127 -0.0153400  0.2444000  1.3310000

 R-square for Endogenous Variables
Alienation67     Anomia67  Powerless67 Alienation71     Anomia71
      0.3212       0.6607       0.6592       0.5763       0.7047
 Powerless71    Education          SEI
      0.6370       0.6936       0.4204

 Parameter Estimates
      Estimate    Std Error   z value    Pr(>|z|)
lam1    0.8885364  0.04150033  21.410348 1.070118e-101
lam2    0.8487223  0.03995708  21.240851 4.005674e-100
lam3    5.3289571  0.42976842  12.399601  2.626270e-35
beta    0.7047276  0.05353754  13.163242  1.428235e-39
gam1   -0.6138170  0.05645164 -10.873326  1.544659e-27
gam2   -0.1741787  0.05391357  -3.230702  1.234864e-03
phi     6.6658511  0.64105526  10.398247  2.525357e-25
psi1    5.3069765  0.47260170  11.229279  2.928570e-29
psi2    3.7412397  0.38756392   9.653220  4.763613e-22
the11   4.0155181  0.34315626  11.701719  1.248976e-31
the22   3.1913382  0.27145244  11.756528  6.536739e-32
the33   3.7010811  0.37341979   9.911315  3.717211e-23
the44   3.6248251  0.29208013  12.410379  2.295635e-35
thd1    2.9441577  0.49980006   5.890671  3.846307e-09
thd2  260.9929854 18.24177314  14.307435  1.966326e-46

lam1  Powerless67 <--- Alienation67
lam2  Powerless71 <--- Alienation71
lam3  SEI <--- SES
beta  Alienation71 <--- Alienation67
gam1  Alienation67 <--- SES
gam2  Alienation71 <--- SES
phi   SES <--> SES
psi1  Alienation67 <--> Alienation67
psi2  Alienation71 <--> Alienation71
the11 Anomia67 <--> Anomia67
the22 Powerless67 <--> Powerless67
the33 Anomia71 <--> Anomia71
the44 Powerless71 <--> Powerless71
thd1  Education <--> Education
thd2  SEI <--> SEI

 Iterations =  85
```



Example where the observed variable is categorical

The code[23] uses the CNES data. At the time of Canadian National Election in 1997, the postal questionnaire was obtained to grasp the citizen's attitude to traditional valuation. The number of respondents was 1529. Variables are shown below.

MBSA2	an ordered factor with levels 'StronglyDisagree', 'Disagree', 'Agree', and 'StronglyAgree', in response to the statement, "We should be more tolerant of people who choose to live according to their own standards, even if they are very different from our own."
MBSA7	an ordered factor with levels 'StronglyDisagree', 'Disagree', 'Agree', and 'StronglyAgree', in response to the statement, "Newer lifestyles are contributing to the breakdown of our society."
MBSA8	an ordered factor with levels 'StronglyDisagree', 'Disagree', 'Agree', and 'StronglyAgree', in response to the statement, "The world is always changing and we should adapt our view of moral behaviour to these changes."
MBSA9	an ordered factor with levels 'StronglyDisagree', 'Disagree', 'Agree', and 'StronglyAgree', in response to the statement, "This country would have many fewer problems if there were more emphasis on traditional family values."

このデータを使ってカテゴリ変数間のポリコリック相関係数を計算させ（ただし\verb!hetcor()!関数に複数の変数を与えた場合、カテゴリ変数同士ではポリコリック相関係数、順序カテゴリと量的変数の間ではポリシリアル相関係数、量的変数同士の間ではピアソンの積率相関係数を自動的に計算してくれる）、

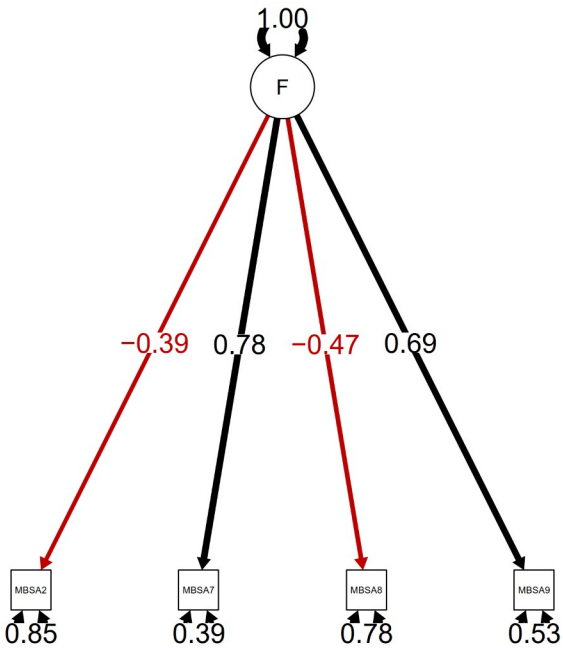
23 https://minato.sip21c.org/advanced-statistics/sem2.R

```
library(sem)
data(CNES)
library(polycor)
print(R.CNES <- hetcor(CNES, std.err=FALSE)$correlations)
model.CNES <- specifyModel(text="
F -> MBSA2, lam1, NA
F -> MBSA7, lam2, NA
F -> MBSA8, lam3, NA
F -> MBSA9, lam4, NA
F <-> F, NA, 1
MBSA2 <-> MBSA2, the1, NA
MBSA7 <-> MBSA7, the2, NA
MBSA8 <-> MBSA8, the3, NA
MBSA9 <-> MBSA9, the4, NA
")
sem.CNES <- sem(model.CNES, R.CNES, N=1529)
summary(sem.CNES, fit.indices=c("GFI","AGFI","CFI","RMSEA"))
library(semPlot)
semPaths(sem.CNES, what="stand", posCol="black", fade=FALSE)
```

The polycoric correlation matrix by **`hetcor()`** is shown below.

```
           MBSA2      MBSA7      MBSA8      MBSA9
MBSA2  1.0000000 -0.3017953  0.2820608 -0.2230010
MBSA7 -0.3017953  1.0000000 -0.3422176  0.5449886
MBSA8  0.2820608 -0.3422176  1.0000000 -0.3206524
MBSA9 -0.2230010  0.5449886 -0.3206524  1.0000000
```

It’s actually CFA, rather than SEM, but AGFI was as high as 0.947.



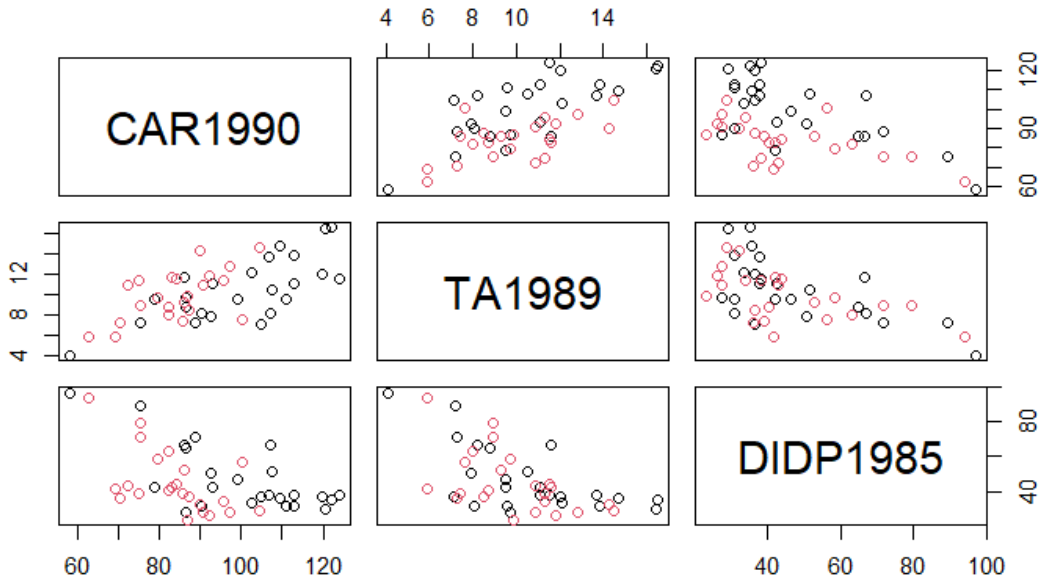

```
 Model Chisquare =  33.2115   Df =  2 Pr(>Chisq) = 6.14066e-08
 Goodness-of-fit index =  0.9893351
 Adjusted goodness-of-fit index =  0.9466755
 RMSEA index =  0.1010603   90% CI: (0.07261014, 0.1326084)
 Bentler CFI =  0.9680971

 Normalized Residuals
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.000003  0.030010  0.207800  0.847900  1.035000  3.830000

 R-square for Endogenous Variables
 MBSA2  MBSA7  MBSA8  MBSA9
0.1516 0.6052 0.2197 0.4717

 Parameter Estimates
     Estimate   Std Error  z value   Pr(>|z|)
lam1 -0.3893289 0.02875484 -13.53959  9.129470e-42 MBSA2 <--- F
lam2  0.7779157 0.02996521  25.96063 1.379394e-148 MBSA7 <--- F
lam3 -0.4686834 0.02839946 -16.50325  3.476850e-61 MBSA8 <--- F
lam4  0.6867992 0.02921502  23.50842 3.344853e-122 MBSA9 <--- F
the1  0.8484230 0.03281417  25.85539 2.116323e-147 MBSA2 <--> MBSA2
the2  0.3948472 0.03567529  11.06781  1.797436e-28 MBSA7 <--> MBSA7
the3  0.7803360 0.03152466  24.75319 2.864281e-135 MBSA8 <--> MBSA8
the4  0.5283069 0.03212698  16.44434  9.208259e-61 MBSA9 <--> MBSA9

 Iterations =  14
```

Applied regression analysis and multilevel model

Multivariate regression model

A regression model that predicts multiple dependent variables using multiple independent variables is called “multivariate regression model”. It's convenient to use **`lavaan`** or **`sem`** to describe structural equations because it allows for the calculation of partial correlations between dependent variables after accounting for the influence of independent variables. As an example, we'll use data for each prefecture including: the **number of cars owned per 100 households in 1990 (CAR1990), the number of traffic accident fatalities per 100,000 population in 1989 (TA1989), the proportion of the population residing in Densely Inhabited Districts (DID) according to the 1985 national census (DIDP1985), prefecture name (PREF), and East/West Japan (REGION). (A REGION value of 1 indicates East Japan, and 2 indicates West Japan). Normally, to examine if there's a difference in traffic accident fatalities, adjusted for car ownership, between East and West Japan, one would use analysis of covariance (ANCOVA). However, the degree of urbanization, represented by the proportion of the population residing in Densely Inhabited Districts, is thought to influence both traffic accident fatalities and car ownership. Additionally, car ownership and traffic accident fatalities are considered to be related. Therefore, multivariate regression analysis can also be performed. First, download the data from internet[24], calculate the correlation matrix, and create a scatter plot matrix.

```
CPA <- read.delim("carpopaccident.txt", stringsAsFactors=TRUE)
cor(CPA[, c("CAR1990", "TA1989", "DIDP1985")])
pairs(CPA[, c("CAR1990", "TA1989", "DIDP1985")], col=CPA$REGION)
```



Using lavaan package, multivariate regression can be done.

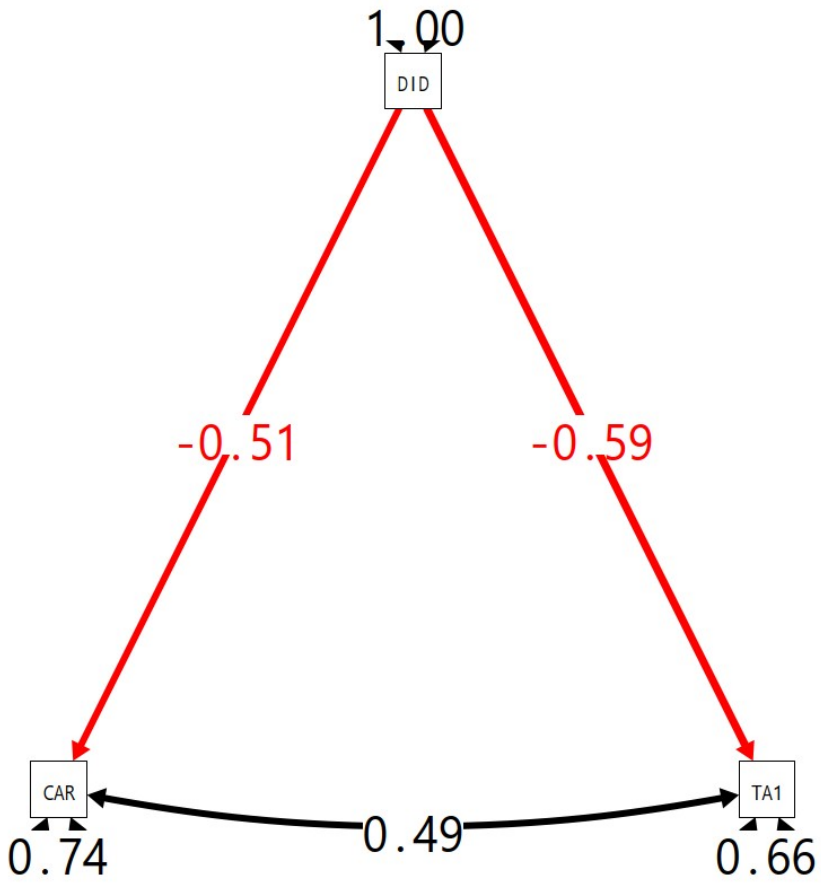
24 https://minato.sip21c.org/advanced-statistics/carpopaccident.txt

```
library(lavaan)
model1 <- 'CAR1990 ~ DIDP1985
TA1989 ~ DIDP1985'
res1 <- sem(model1, data=CPA[, c("CAR1990", "TA1989", "DIDP1985")])
summary(res1, standardized=TRUE, fit.measures=TRUE)
library(semPlot)
semPaths(res1, what="stand", posCol="black", negCol="red", fade=FALSE,
 edge.label.cex=2)
```

The results are shown below.

```
lavaan (0.5-20) converged normally after  28 iterations
  Number of observations                            47
  Estimator                                         ML
  Minimum Function Test Statistic                0.000
  Degrees of freedom                                 0

Model test baseline model:
  Minimum Function Test Statistic               47.096
  Degrees of freedom                                 3
  P-value                                        0.000

User model versus baseline model:
  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.000

Loglikelihood and Information Criteria:
  Loglikelihood user model (H0)               -489.488
  Loglikelihood unrestricted model (H1)       -489.488
  Number of free parameters                          5
  Akaike (AIC)                                 988.975
  Bayesian (BIC)                               998.226
  Sample-size adjusted Bayesian (BIC)          982.544

Root Mean Square Error of Approximation:
  RMSEA                                          0.000
  90 Percent Confidence Interval          0.000  0.000
  P-value RMSEA <= 0.05                          1.000

Standardized Root Mean Square Residual:
  SRMR                                           0.000

Parameter Estimates:
  Information                                 Expected
  Standard Errors                             Standard

Regressions:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
  CAR1990 ~
    DIDP1985         -0.441    0.108   -4.082    0.000   -0.441   -0.512
  TA1989 ~
    DIDP1985         -0.086    0.017   -4.947    0.000   -0.086   -0.585

Covariances:
```

```
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
  CAR1990 ~~
    TA1989           14.668    4.832    3.035    0.002   14.668    0.494

Variances:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
    CAR1990         183.639   37.882    4.848    0.000  183.639    0.738
    TA1989            4.805    0.991    4.848    0.000    4.805    0.658
```

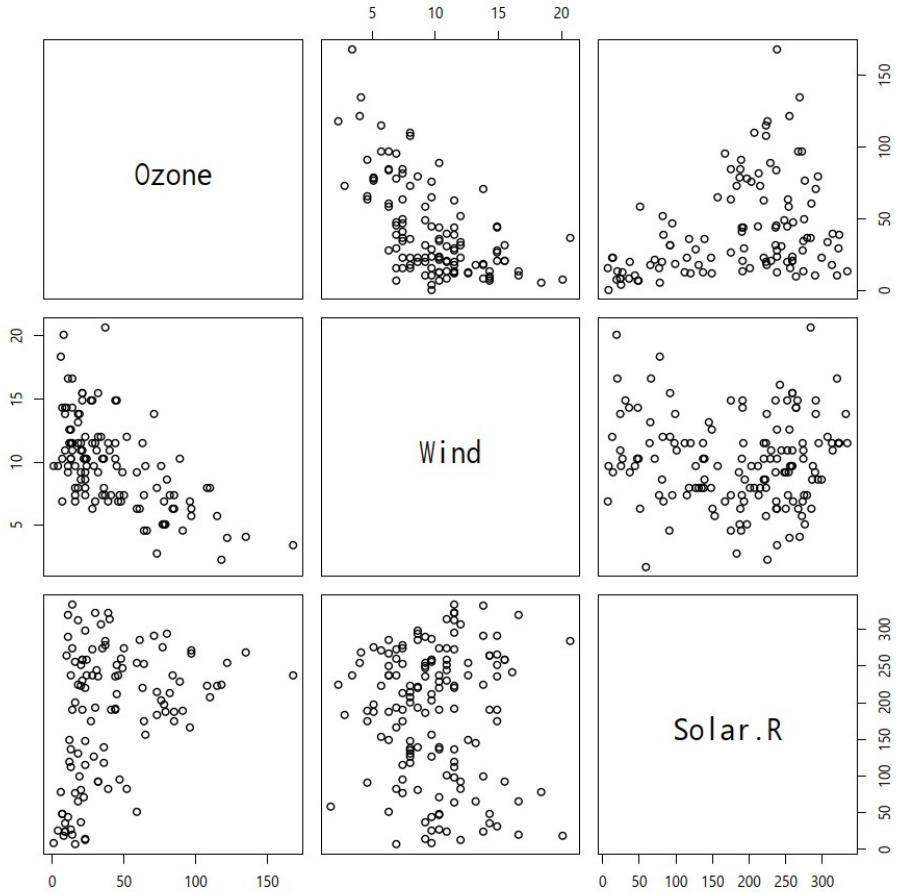


Nonlinear regression model

In R, you can easily fit data to models: linear regression using the **`lm()`** function, and generalized linear models—which include logistic regression (where the dependent variable follows a binomial distribution) and Poisson regression (where it follows a Poisson distribution)—using the **`glm()`** function. Nonlinear regression analysis is used when the independent variables are not linearly combined. For simple continuous functions, you can use the **`nls()`** function. More generally, the **`optim()`** function allows you to estimate parameters that fit any function shape to your data.

You can use the following code to regress ozone concentration on an exponential function of wind speed and a linear combination of solar radiation using the **airquality** dataset, which is built into R and contains New York air quality data. Finally, **predict()** is used to display the ozone concentrations predicted by this nonlinear regression model when solar radiation is at its mean, and wind speeds are 0, 5, 10, 15, 20, and 25 meters per second[25].

```
data(airquality)
pairs(airquality[, c("Ozone", "Wind", "Solar.R")])
resaq <- lm(Ozone ~ Solar.R + Wind, data=airquality)
summary(resaq)
AIC(resaq)
predict(resaq, list(Wind=0:5*5,
Solar.R=rep(mean(resaq$model$Solar.R), 6)))
AQ <- subset(airquality, !is.na(Ozone)&!is.na(Solar.R)&!is.na(Wind))
resmr <- nls(Ozone ~ a*exp(-b*Wind) + c*Solar.R,
             start=list(a=200, b=0.2, c=1), data=AQ)
summary(resmr)
AIC(resmr)
predict(resmr, list(Wind=0:5*5, Solar.R=rep(mean(AQ$Solar.R), 6)))
```



25 https://minato.sip21c.org/advanced-statistics/nls.R

The result from lm() follows.

```
Call:
lm(formula = Ozone ~ Solar.R + Wind, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-45.651 -18.164  -5.959  18.514  85.237

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 77.24604    9.06751   8.519 1.05e-13 ***
Solar.R      0.10035    0.02628   3.819 0.000224 ***
Wind        -5.40180    0.67324  -8.024 1.34e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.92 on 108 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.4495,    Adjusted R-squared:  0.4393
F-statistic: 44.09 on 2 and 108 DF,  p-value: 1.003e-14

> AIC(resaq)
[1] 1033.816

> predict(resaq, list(Wind=0:5*5,
Solar.R=rep(mean(resaq$model$Solar.R), 6)))
        1         2         3         4         5         6
 95.79102  68.78203  41.77304  14.76406 -12.24493 -39.25391
```

The results from nonliear regression model are shown below. AIC becomes smaller than lm() and the predicted values are not negative.

```
Formula: Ozone ~ a * exp(-b * Wind) + c * Solar.R

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 215.42457   33.11390   6.506 2.49e-09 ***
b   0.24432    0.03331   7.335 4.32e-11 ***
c   0.08639    0.02014   4.290 3.90e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.01 on 108 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 9.951e-06

> AIC(resmr)
```

```
[1] 1006.24
> predict(resmr, list(Wind=0:5*5, Solar.R=rep(mean(AQ$Solar.R), 6)))
[1] 231.38999  79.46401  34.68228  21.48240  17.59160  16.44475
```

Analysis of dose-response relation

急性毒性試験でよく使われる用量反応関係も非線形回帰の一種なので、簡単に説明しておく。毒物を実験動物に投与した場合、用量（dose）や血中濃度に応じて標的臓器や個体の反応程度が変化するのだが、有害物の負荷量としての投与量（dose）に対する反応割合（＝反応した個体数／その dose を受けた総個体数）との関係を集団レベルでみると、S 字曲線になることが多い。原因は、反応（感受性）に個体差があることで、通常、累積対数正規分布で近似される。半数の個体が反応を示す負荷量を半数影響量(ED50)と呼ぶ。急性毒性試験では半数致死量(LD50)が良く使われ、推定にはプロビット分析\footnote{プロビット分析では、$$F^*(X_i)=\Phi(\beta_0+\beta_1 X_i), \Phi(z)=\int^z_{-\infty} {1\over \sqrt{2\pi}} e^{-x^2/2} dx$$というモデルを当てはめる。}やロジット分析\footnote{ロジット分析では、対数オッズが用量の線形関数となり、$$F^*(X_i)=\Lambda(\beta_0 + \beta_1 X_i), \Lambda(z)={e^z \over 1+e^z}$$というモデルを当てはめる。}が使われる。

用量反応関係のモデルとしては、ワンヒットモデル\footnote{発がん物質が 1 回遺伝子に衝突し損傷を与えると、その細胞ががん化するというモデル。曝露量 D に対して細胞ががん化する確率 P(D)は、1 から 1 回も衝突しない確率を引いた値として得られ、$$\hbox{P(D)}=1-\hbox{h(0)}=1-e^{-\hbox{qD}}$$とモデル化される。低用量域では発がんリスクが用量に比例するので、この比例定数 q を発がんスロープファクタと呼ぶ。}や、線型多段階モデル\footnote{1 つの細胞ががん化するために k 段階の反応が一定の順序で起こる必要があり、各段階の反応率が用量の一次式で表されると考えるモデル。$$\hbox{P(D)}=1-e^{-(q_0+q_1D+q_2D^2+...+q_kD^k)}$$と定式化される。米国 EPA が用いているのは、このモデルに$q_1>0$という制約をつけた Crump のモデルである。通常は、P(0)によるリスクを除いた曝露量 D での発がんリスク、つまり過剰リスク$$\hbox{R(D)}={\hbox{P(D)}-\hbox{P(0)} \over 1-\hbox{P(0)}}$$を考える。このモデルでは D が 0 に近いとき R(D)は近似的に$q_1\hbox{D}$となるので、低用量域では過剰リスクが用量に比例する。}のようなものもよく使われている。

R を使っていると、このように限定された目的には、専用のパッケージが存在することが多く、それを使う方が\verb!nls()!関数で頑張るより便利である。ここでは\verb!drc!パッケージの使い方を示す。サンプルデータとして、\verb!doBy!パッケージに入っている青虫（\verb!budworm!）のデータを使ってみる（データの出典は、Venables and Ripley (1999) {\it Modern Applied Statistics with S-Plus.\/}, Springer である）。{\it trans}-cypermethrin という殺虫剤の用量（μg単位）を何段階かで変えて投与したときの、雄と雌の青虫（タバコの葉を食べる蛾の幼虫）の反応を見たデータである。変数としては、\verb!sex!、\verb!dose!、\verb!ndead!、\verb!ntotal!が含まれている。つまり、1 行が 1 個体ではなく、dose ごとに雄雌それぞれ 1 行が与えられており、その処理の青虫の個体数が\verb!ntotal!、そのうち死亡した個体数が\verb!ndead!に与えられている。以下のように呼び出す。

\begin{itembox}[l]{https://minato.sip21c.org/advanecd-statistics/dr.R(1)}
\small\begin{verbatim}
if (require(doBy)==FALSE) {
 install.packages("doBy"); library(doBy) }
data(budworm)
\end{verbatim}\end{itembox}

一般化線型モデルの関数\verb!glm()!を使ってロジスティック回帰分析し、それを\verb!dose.LD50()!関数に与えるという方法で分析できる（2007 年頃の\verb!doBy!パッケージにはこの関数が含まれていた。ただし、現在 cran からインストールできる\verb!doBy!パッケージでは\verb!dose.LD50()!関数が無いので、まずそれを定義する（下枠内はかつて存在したコードから作成したものである）。現在の\verb!doBy!パッケージは多種多様な方法でのグループ統計量の処理をする方向に特化し、例えば、修正平均の計算\footnote{\url{https://mran.microsoft.com/web/packages/doBy/vignettes/LSmeans.pdf}}をする場合などに便利である。

\begin{itembox}[l]{https://minato.sip21c.org/advanecd-statistics/dr.R(2)}
\scriptsize\begin{verbatim}
```
.ratioVar <- function(x, num, den, numval){
  m1 <- x
  beta <- coef(m1)
  numvec <- rep(0,length(beta))
  denvec <- rep(0,length(beta))
  numvec[num] <- numval
  denvec[den] <- 1
  M <- rbind(numvec, denvec)
  vcv <- summary(m1)$cov.scale
  beta2 <- M %*% beta
  vcv2  <- M %*% vcv %*% t(M)
  muvec <- c(1/beta2[2], -beta2[1]/(beta2[2]^2))
  ratiovar <- t(muvec) %*% vcv2 %*% muvec
  return(ratiovar)
}
.ratio <- function(x, num, den, numval, sign=-1){
  m1 <- x
  beta <- coef(m1)
  numvec <- rep(0, length(beta))
  denvec <- rep(0, length(beta))
  numvec[num] <- numval
  denvec[den] <- 1
  M <- rbind(numvec, denvec)
  beta2 <- M %*% beta
  ratio <- sign*beta2[1, 1]/beta2[2, 1]
  return(ratio)
}
.ld50 <- function(x, num, den, numval){
 est  <- .ratio(x, num, den, numval)
 vare <- .ratioVar(x, num, den, numval)
 result <- c("ld50"=est, lower=est-1.96*sqrt(vare), upper=est+1.96*sqrt(vare))
 return(result)
}
dose.LD50 <- function (x, lambda) {
 if (length(which(is.na(lambda))) != 1) {
  stop("lambda must contain exactly one entry which is NA") } else {
  den <- which(is.na(lambda))
```

```
  num <- which(!is.na(lambda))
  numval <- lambda[num]
  value <- .ld50(x, num, den, numval)
  return(value) }
}
```

このように定義した\verb!dose.LD50()!関数を使えば、\verb!glm()!の結果から LD50 を計算することができる。

```
mx <- glm(ndead/ntotal ~ sex + dose, weights=ntotal,
 data=budworm, family=binomial)
summary(mx)
dose.LD50(mx, c(1, 1, NA)) # for males
dose.LD50(mx, c(1, 0, NA)) # for females
```

結果は以下のように得られる。

```
> summary(mx)
Call:
glm(formula = ndead/ntotal ~ sex + dose, family = binomial,
    data = budworm, weights = ntotal)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5566  -1.3326   0.3384   1.1254   1.8838

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.13462    0.32029  -6.665 2.66e-11 ***
sexmale      0.96855    0.32954   2.939  0.00329 **
dose         0.15996    0.02341   6.832 8.39e-12 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:  27.968  on  9  degrees of freedom
AIC: 64.078
> dose.LD50(mx, c(1, 1, NA)) # for males
ld50.numvec       lower       upper
   7.289943    4.620782    9.959105
```

```
> dose.LD50(mx, c(1, 0, NA)) # for females
ld50.numvec       lower       upper
  13.34505    10.16512    16.52498
\end{verbatim}\end{screen}
```

5\%水準で有意な性差があることもわかる。次に性差を無視して dose の対数で回帰してみる。結果としての LD50 を見るときに指数をとらなくてはいけないことに注意。

```
\begin{itembox}[l]{https://minato.sip21c.org/advanecd-statistics/dr.R(4)}
\small\begin{verbatim}
m2 <- glm(ndead/ntotal ~ log(dose), weights=ntotal,
 data=budworm, family=binomial)
summary(m2)
exp(dose.LD50(m2, c(1, NA))) # same results as drm of drc
\end{verbatim}\end{itembox}
```

結果は以下。

```
\begin{screen}\small\begin{verbatim}
> summary(m2)
Call:
glm(formula = ndead/ntotal ~ log(dose), family = binomial,
    data = budworm, weights = ntotal)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7989  -0.8267  -0.1871   0.8950   1.9850
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7661     0.3701  -7.473 7.82e-14 ***
log(dose)     1.4525     0.1783   8.147 3.74e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:  16.984  on 10  degrees of freedom
AIC: 51.094
Number of Fisher Scoring iterations: 4
> exp(dose.LD50(m2, c(1, NA))) # same results as drm of drc
ld50.numvec       lower       upper
  6.714995    5.362009    8.409377
\end{verbatim}\end{screen}
```

しかし、こんなに面倒なことをしなくても、\verb!drc!パッケージを使うときわめて簡単である。

```
\begin{itembox}[l]{https://minato.sip21c.org/advanecd-statistics/dr.R(5)}
\small\begin{verbatim}
if (require(drc)==FALSE) { install.packages("drc"); library(drc) }
```

```
m3 <- drm(ndead/ntotal ~ dose, weights=ntotal,
 data=budworm, fct=LL.2(), type="binomial") # LL.2 is log-logistic model
summary(m3)
ED(m3, 50, interval="delta")
plot(m3)
```

以下の結果が得られる。dose を対数変換して\verb!glm()!に与えた結果を\verb!dose.LD50()!に与えた結果とほぼ一致している。結果オブジェクトを\verb!plot()!に渡すだけでグラフも描ける。

```
> summary(m3)
Model fitted: Log-logistic (ED50 as parameter) with
 lower limit at 0 and upper limit at 1 (2 parms)
Parameter estimates:
             Estimate Std. Error  t-value p-value
b:(Intercept) -1.45252    0.17830 -8.14645       0
e:(Intercept)  6.71483    0.77084  8.71101       0
> ED(m3, 50, interval="delta")
Estimated effective doses
(Delta method-based confidence interval(s))
     Estimate Std. Error   Lower  Upper
1:50  6.71483    0.77084 5.20400 8.2257
```

$$\includegraphics[width=8cm]{drcbudworm1.pdf}$$

\verb!drc!パッケージで性差を考慮するには、\verb!curveid=!オプションで指定するだけである。

```
m4 <- drm(ndead/ntotal ~ dose, curveid=sex, weights=ntotal,
         data=budworm, fct=LL.2(), type="binomial")
summary(m4)
ED(m4, 50, interval="delta")
plot(m4)
```

結果は以下の通り得られる。LD50 は雌が$9.9\pm1.8 \mu g$、雄が$4.7\pm 0.7 \mu g$（$\pm$の後の値は標準誤差）と推定される。

```
> summary(m4)
Model fitted: Log-logistic (ED50 as parameter) with
 lower limit at 0 and upper limit at 1 (2 parms)
Parameter estimates:
         Estimate Std. Error  t-value p-value
b:male   -1.81592    0.30588 -5.93675       0
```

```
b:female -1.30705    0.24102 -5.42311       0
e:male    4.72170    0.66779  7.07060       0
e:female  9.87097    1.78005  5.54534       0
> ED(m4, 50, interval="delta")
Estimated effective doses
(Delta method-based confidence interval(s))
         Estimate Std. Error   Lower   Upper
female:50  9.87097    1.78005 6.38214 13.3598
male:50    4.72170    0.66779 3.41285  6.0306
\end{verbatim}\end{screen}
```

$$\includegraphics[width=8cm]{drcbudworm2.pdf}$$

次に、官庁のウェブサイトで多くの毒性試験データが公開されているので、\url{https://www.wam.go.jp/wamappl/bb11gs20.nsf/0/f8e1fb7d07413544492571d1000bff64/$FILE/20060818siryou3-3_1_6.pdf}からデータを取り出す\footnote{元々は\url{https://dra4.nihs.go.jp/mhlw_data/home/paper/paper583-39-1a.html}にあった。一時リンクが切れていたが復活したようだ。}。ラットを使って行われた 2-mercaptobenzimidazole の急性毒性試験である。報告書ではプロビット分析または Behrens-Karber 法によって、LD50 を雌で 208、雄で 218 と示されている。上の例と同じく対数ロジット法（\verb!drc!パッケージで\verb!fct=LL.2()!オプションを使う）で分析するコードを示す。なお、\verb!drc!パッケージで\verb!fct!オプションに与えることができる関数の一覧は、\verb!drc!パッケージをロードした状態で、\verb!getMeanFunctions()!と打てば得られる。応答変数のタイプも\verb!type!オプションに\verb!"continuous"!や\verb!"binomial"!、\verb!"Poisson"!などの文字列を与えることで指定可能である。

```
\begin{itembox}[l]{https://minato.sip21c.org/envhlth/dr2.R}\small\begin{verbatim}
if (require(drc)==FALSE) { install.packages("drc"); library(drc) }
rats <- data.frame(
 sex = factor(c(rep(1, 7), rep(2, 7)), labels=c("M", "F")),
 dose = rep(c(0, 79, 119, 178, 267, 400, 600), 2),
 ndead = c(0, 0, 0, 1, 4, 5, 5, 0, 0, 0, 1, 5, 5, 5),
 ntotal = rep(5, 14))
mx <- drm(ndead/ntotal ~ dose, curveid=sex, weights=ntotal,
 data=rats, fct=LL.2(), type="binomial")
summary(mx)
ED(mx, 50, interval="delta")
plot(mx)
\end{verbatim}\end{itembox}
```

結果は次のように得られる。

```
\begin{screen}\small\begin{verbatim}
> summary(mx)
Model fitted: Log-logistic (ED50 as parameter) with
 lower limit at 0 and upper limit at 1 (2 parms)
Parameter estimates:
    Estimate Std. Error   t-value p-value
b:M  -7.77184    3.24901  -2.39206  0.0168
```

b:F -29.25825 198.69956 -0.14725 0.8829
e:M 218.01305   22.80926   9.55809 0.0000
e:F 186.63562   60.37334   3.09136 0.0020
> ED(mx, 50, interval="delta")
Estimated effective doses
(Delta method-based confidence interval(s))
    Estimate Std. Error   Lower Upper
F:50 186.636   60.373 68.306 304.97
M:50 218.013   22.809 173.308 262.72
\end{verbatim}\end{screen}

$$\includegraphics[width=8cm]{drcrats.pdf}$$

Multilevel model

Many tutorial texts are available in English[26][27], but this explanation is based on the textbook in Japanese (Fujino Y, Kondo N, Takeuchi F (2013) The multilevel analysis for health and medical professionals. Shindan-to-chiryo, ISBN:978-4-7878-2053-2), since the data can be downloaded from the publisher's web site[28].

In the multilevel model, the dependent variable obtained for each individual is explained by the combination of the variables obtained for each individual and the variables obtained at group (meso- or macro-) level. Usual regression model can also treat the group level independent variables as fixed effect of the common factor similarly affecting different individuals, but multilevel model considers such group level variables (variate) as values sampled from a kind of distribution, and thus the effects of such variables are called as random effects.

The fixed effect means that how large the effect is among the different levels of group variable, but expected value of random effect is always zero, the important information from the random effect is the extent of variation. In addition, within the subgroup for each level in the group level variable, all individual level variables are expected to be correlated, which is called as intra-class correlation (ICC).

Obviously, repeated measures data can be considered in this framework, if we consider the inter-individual difference as random effect. According to the textbook (Fujino et al., ibid.), multilevel analysis has the following 3 purposes.
(1) Considering hierarchy in analysis
(2) Examining the effect of macro-level variable
(3) Examining the variation among the levels in group level variable, whether the variation is attributable to individual characteristics or group characteristics

The sample size calculation for multilevel analysis is not supported in EZR, PS or jamovi (though the Pamlj module can support the sample size calculation for SEM). G*Power has the menu to calculate sample size for repeated measures ANOVA, but in general, special tools such as MLPowSim[29]

26 https://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf
27 https://quantdev.ssri.psu.edu/tutorials/r-bootcamp-introduction-multilevel-model-and-interactions
28 https://www.shindan.co.jp/np/isbn/9784787820532/
29 https://www.bristol.ac.uk/cmm/software/mlpowsim/

(developed by Bristol University) or PinT[30] (by Prof. Tom Snijders in Oxford Univ.) are needed. However, as an alternative, sum of the sample sizes for each strata by the levels of group level variable, calculated by the equations for usual linear regression models can be used.

The essence of multilevel analysis

マルチレベル分析は、数学的には混合効果モデルの枠組みで扱うことができる。繰り返しになるが、従属変数への効果を固定効果(Fixed Effects)と変量効果(Random Effects)の混合と捉えることで、個体レベルの変数の効果を固定効果として、集団レベルのばらつき（言い換えると、集団ごとに固定効果の傾きと切片がランダムに異なること）を変量効果として分析するわけである。

R の\verb!lme4!パッケージでは、線形混合効果モデルのみでなく、従属変数が正規分布に従わない一般化線形混合効果モデルや、非線形混合効果モデルも分析することが可能とされているが、ここでは線形混合効果モデルのみ説明する。

モデルの指定の方法は、

\begin{screen}\small\begin{verbatim}
lmer(resp ~ FEexpr + (REexpr1 | factor1) + (REexpr2 | factor2) + ..., data=df)
\end{verbatim}\end{screen}

のようにする。ここで、\verb!resp!は従属変数（応答変数）、\verb!FEexpr!は固定効果の変数、\verb!(REexpr1 | factor1)!と\verb!(REexpr2 | factor2)!は変量効果の変数項または共分散因子の構造を示す項である。変量効果の変数項の数はいくらでも多くとれるが、通常は少数にとどめる。右辺の書き方を下表にまとめる（\verb!g!、\verb!g1!、\verb!g2!は変量効果をみたいグループレベル変数、\verb!x!は固定効果をみたい共変量、\verb!o!は既知のオフセットである）。

$$\vbox{\small
\begin{tabular}{lp{8cm}}
\hline
式 & 意味\cr
\hline
\verb!(1 | g)! & \verb!g!の各グループごとにランダム切片がある。これらの切片の平均と標準偏差が推定される。\cr
\verb!0 + offset(o) + (1 | g)! & \verb!g!の各グループごとにランダム切片があり、ゼロでない変量効果の切片が既知のオフセット値\verb!o!。\cr
\verb!(1 | g1/g2)! & \verb!g1!内にネストされた\verb!g2!があるとき。\cr
\verb!(1 | g1) + (1 | g2)! & ネストしていない独立なグルーピングとして\verb!g1!と\verb!g2!があり、それらの変量効果をみたいとき。\cr
\verb!x + (x | g)! & \verb!x!の固定効果があり、\verb!g!のグループごとに\verb!x!から応答変数への効果の切片と傾きの両方が変動するとき。\cr
\verb!x + (x || g)! & デフォルトでは同一の変量効果項内の全変数は相関していると仮定するが、こう書けば無相関な切片と傾きを仮定できる。\cr
\hline
\end{tabular}

30 https://www.stats.ox.ac.uk/~snijders/multilevel.htm

}$$

なお、モデルの当てはめのとき、とくにオプションを指定しない限り、REML（制限付き最尤法）が用いられる。REML は、普通の最尤法では推定値にバイアスがかかるという問題への対処として提案された方法の 1 つである。REML では、固定効果を除去するデータの線形結合を考慮することによって変量効果を推定する。このため、固定効果だけが異なる 2 つのネストされたモデル（片方がもう片方を含んでいるようなモデル）を尤度比検定で評価したい場合は、REML は使えない。REML ではなく普通の最尤法(ML)を使いたい場合は、オプションとして\verb!REML=FALSE!をつければ良い。

Example 1: Intervention study in multiple institutions

藤野ら(2013)の pp.69-70 に示されている多施設介入試験の分析を R の\verb!lmer()!関数で実行する方法を示す。\verb!lmer()!関数は\verb!lme4!パッケージに含まれているので、予め\verb!install.packages("lme4", dep=TRUE)!でインストールしておく必要がある\footnote{後述するように、筆者は\verb!lmerTest!パッケージを使う方が良いと思うが、思想の問題なので何とも言えない。}。実行コードは以下の通り。

\begin{screen}\small\begin{verbatim}
x <- read.csv("https://www.shindan.co.jp/np/filedata/00205300_4.csv")
library(lme4)
res <- lmer(cholesterol ~ cholesterol_base + intervention + (1 | clinic),
 data=x, REML=FALSE)
summary(res)
confint(res)
\end{verbatim}\end{screen}

Stata の出力である図 9-5 の数値と比べると、Wald 検定とのその p 値、固定効果（介入とベースラインのコレステロール値と切片）の係数の z と p 値を除けば求めることができている。固定効果の係数については、z の代わりに t value が提示されている。{\bf p 値は敢えて表示していない}とのことである。自由度が簡単には求められないからというのが大きな理由である。

それでも p 値が欲しい場合は、\url{https://mindingthebrain.blogspot.jp/2014/02/three-ways-to-get-parameter-specific-p.html}に解説されているように、自由度が十分に大きい t 分布は正規分布とほぼ同じだからと考えれば、

\begin{screen}\begin{verbatim}
1-pnorm(summary(res)$coefficients[,"t value"])
\end{verbatim}\end{screen}

で p 値を計算してしまうこともできる。また、\verb!doBy!パッケージの\verb!LSmeans()!関数を使い、\verb!adjust.df=FALSE!オプションを指定すると、自由度がサンプルサイズより 1 つ小さい t 分布を使って p 値を計算することができるようである（注：未確認である）。

{\bf SAS や Stata で得られるのと同等な p 値を計算する}ためには、

\begin{screen}\small\begin{verbatim}

```
install.packages("lmerTest", dep=TRUE)
\end{verbatim}\end{screen}
```

によって\verb!lmerTest!パッケージをインストールしておけば、

```
\begin{screen}\small\begin{verbatim}
x <- read.csv("https://www.shindan.co.jp/np/filedata/00205300_4.csv")
library(lmerTest)
res <- lmer(cholesterol ~ cholesterol_base + intervention + (1 | clinic),
 data=x, REML=FALSE)
summary(res)
confint(res)
\end{verbatim}\end{screen}
```

のように、まったく同じ関数指定でも p 値が得られる\footnote{SAS の\verb!PROC MIXED!同様、Satterthwaite の近似で自由度と p 値を計算してくれると書かれている。}。あるいは、\verb!pbkrtest!パッケージを使えば Kenward-Roger の近似で自由度を出すことができるので、t 分布の累積確率密度関数を使って p 値を出すこともできる。

実は既出の\verb!doBy!パッケージを使うと、\verb!lme4!パッケージの\verb!lmer()!関数の結果を\verb!LSmeans()!関数に渡すだけで（ただし\verb!effect=!オプションを適切に指定する必要がある）、Kenward-Roger の近似自由度を使った p 値を計算してくれるようである。

以下、この例で、データの性状を見るための作図からマルチレベル分析まで、一通りの操作をするコードをまとめて示す。

```
\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/multilev.R(1)}
\small\begin{verbatim}
x <- read.csv("https://www.shindan.co.jp/np/filedata/00205300_4.csv")
# graph1
layout(matrix(1:4, 2))
hist(x$cholesterol)
hist(x$cholesterol_base)
matplot(rbind(x$cholesterol_base, x$cholesterol), type="l",
 col=x$intervention+2, lty=1, lwd=1, axes=FALSE)
axis(1, 1:2, c("base","after"))
axis(2, 3:7*50)
stripchart(cholesterol-cholesterol_base ~ intervention, data=x,
 method="stack", ylab="intervention")
\end{verbatim}\end{itembox}
```

データの性状を見るためのグラフは次のように描かれる。

$$\includegraphics[width=9cm]{multilevel-graph1.pdf}$$

分析は、まず介入の有無も施設の違いも無視して、前後でのコレステロール値の差があるかどうかだけ、対応のある t 検定で調べてみると、

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/multilev.R(2)}
\small\begin{verbatim}
t-test
t.test(x$cholesterol_base, x$cholesterol,
 paired=TRUE, var.equal=FALSE)
# graph2
layout(1)
plot(cholesterol ~ cholesterol_base, pch=intervention+1, data=x)
legend("topleft", pch=1:2,
 legend=c("without intervention","with intervention"))
\end{verbatim}\end{itembox}

\begin{screen}\small\begin{verbatim}
        Paired t-test

data: x$cholesterol_base and x$cholesterol
t = 8.0383, df = 275, p-value = 2.713e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 10.96801 18.08272
sample estimates:
mean of the differences
            14.52536
\end{verbatim}\end{screen}

と、有意水準 5\%で統計学的に有意な差があるといえる。

$$\includegraphics[width=8cm]{multilevel-graph2.pdf}$$

次に、藤野ら(2013)p.69 図 9-4 に示されている Stata による回帰分析の結果をと比較するため、共分散分析で、介入の有無とベースラインのコレステロール値の、治療後のコレステロール値への交互作用効果を調べてみる。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/multilev.R(3)}
\small\begin{verbatim}
ANCOVA
x$interventionF <- as.factor(x$intervention)
res1 <- lm(cholesterol ~ cholesterol_base * interventionF, data=x)
summary(res1)
\end{verbatim}\end{itembox}

\begin{screen}\small\begin{verbatim}
Call:
lm(formula = cholesterol ~ cholesterol_base * interventionF,
    data = x)

```
Residuals:
    Min      1Q  Median      3Q     Max
-52.619 -18.290  -2.475  16.547  87.592

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   41.05775   12.48988   3.287  0.00114 **
cholesterol_base               0.78941    0.05094  15.496  < 2e-16 ***
interventionF1               111.80484   24.42985   4.577 7.19e-06 ***
cholesterol_base:interventionF1  -0.48289    0.09812  -4.922 1.49e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.97 on 272 degrees of freedom
Multiple R-squared:  0.483,   Adjusted R-squared:  0.4773
F-statistic: 84.72 on 3 and 272 DF,  p-value: < 2.2e-16
\end{verbatim}\end{screen}
```

と、交互作用効果（\verb!cholesterol_base:interventionF1!の行）も有意水準 5\%で統計学的に有意であった。つまり、介入した群としなかった群の間で、ベースラインのコレステロール値と治療後のコレステロール値の関係が有意に異なっていた。もし解析の目的がベースラインコレステロール値と治療後コレステロール値の関係を調べることであれば、共分散分析によって修正平均を比較するよりも、2 群別々に分析すべきということになるのだが、ここでの分析の目的は介入効果の評価なので、その方法では不適切である。

そこで、前掲書図 9-4 と同じ結果を得るには交互作用効果を入れずに lm()関数で線形回帰すれば良い。

```
\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/multilev.R(4)}
\small\begin{verbatim}
res2 <- lm(cholesterol ~ cholesterol_base + interventionF, data=x)
summary(res2)
# graph3
plot(cholesterol ~ cholesterol_base, pch=intervention+1, data=x,
 col=topo.colors(10)[clinic])
legend("topleft", pch=1:2,
 legend=c("without intervention","with intervention"))
legend("bottomright", pch=2, col=topo.colors(10),
 legend=1:10, title="clinic")
\end{verbatim}\end{itembox}
```

```
\begin{screen}\small\begin{verbatim}
Call:
lm(formula = cholesterol ~ cholesterol_base + interventionF,
    data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-51.755 -20.040  -2.563  16.194  91.172
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      72.49400   11.18021   6.484 4.16e-10 ***
cholesterol_base  0.65923    0.04535  14.536  < 2e-16 ***
interventionF1   -7.42741    3.27763  -2.266   0.0242 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.05 on 273 degrees of freedom
Multiple R-squared:  0.437,   Adjusted R-squared:  0.4329
F-statistic:   106 on 2 and 273 DF,  p-value: < 2.2e-16
\end{verbatim}\end{screen}
```

のようになり、前掲書図 9-4 と同じ結果が得られる。このとき、ベースラインのコレステロール濃度の影響を調整した上で、有意な介入効果が得られていると考えられる。

$$\includegraphics[width=10cm]{multilevel-graph3.pdf}$$

しかし施設ごとに色を変えてプロットし直してみると、施設の効果がありそうに思えるので、これはマルチレベル分析にすべきである。\verb!lmerTest!パッケージの\verb!lmer()!関数により得られた結果オブジェクトを\verb!summary()!に渡せば、

```
\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/multilev.R(5)}
\small\begin{verbatim}
# multilevel
library(lmerTest)
res3 <- lmer(cholesterol ~ cholesterol_base + intervention +
 (1 | clinic), data=x, REML=FALSE)
summary(res3)
confint(res3)
\end{verbatim}\end{itembox}
```

により、以下の結果が得られる。

```
\begin{screen}\small\begin{verbatim}
Linear mixed model fit by maximum likelihood t-tests use Satterthwaite
  approximations to degrees of freedom [lmerMod]
Formula: cholesterol ~ cholesterol_base + intervention + (1 | clinic)
   Data: x

     AIC      BIC   logLik deviance df.resid
  2594.9   2613.0  -1292.4   2584.9      271

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8741 -0.7486 -0.1038  0.5595  3.0095
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 clinic   (Intercept) 141.7    11.90
 Residual             637.0    25.24
Number of obs: 276, groups:  clinic, 10

Fixed effects:
                 Estimate Std. Error        df t value Pr(>|t|)
(Intercept)      116.72601   13.27558 163.47000   8.793    2e-15 ***
cholesterol_base   0.47132    0.05239 247.26000   8.997   <2e-16 ***
intervention      -1.74572    3.26249 275.75000  -0.535    0.593
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) chlst_
cholstrl_bs -0.945
interventin  0.120 -0.241
\end{verbatim}\end{screen}
```

係数の信頼区間は、同じ結果オブジェクトを\verb!confint()!関数に渡せば得られる。\verb!lme4!の開発者による解説文書には、Wald の近似を使い、ブートストラップ法で推定した値が表示されると書かれている（だから\verb!lme4!パッケージでは p 値を表示しないのである）。以上の手順で、Stata で得られる結果を、ほぼすべて R でも得ることができる。マルチレベル分析にすると、回帰分析では有意であった intervention の効果が有意でない、というのがポイントである。この場合、介入効果があったように見えていたのは、施設間差による artefact であったと考えられる。

```
\begin{screen}\small\begin{verbatim}
Computing profile confidence intervals ...
                   2.5 %      97.5 %
.sig01           6.4527670  21.7732086
.sigma          23.2278626  27.5583053
(Intercept)     87.9180189 145.5631716
cholesterol_base  0.3581452   0.5886708
intervention     -8.4035413   4.8428531
\end{verbatim}\end{screen}
```

Example 2: Animal experiments to consider inter-individual variation

藤野(2013)pp.73-74 に書かれている、遺伝子導入した豚のデータを R で分析してみる。このデータはマクロレベル変数の影響を調べる目的で提示されている。A, B の異なる遺伝子のどちらかを導入した豚と、遺伝子導入していないコントロール豚の 3 群（各群 5 頭）を対象に、心臓に高頻度ペーシング\footnote{電気刺激によって心臓の動く回数を増加させる方法のこと。}を 7 日間行い、心房細動が誘発される頻度が遺伝子導入と関連があるかを調べたというデザインである。毎日 2 分間心電図を記録し、出現した波形のうち、心房細動（P 波がないパタン）が出現する割合を調べている。

変数は、\verb!n_obs!が出現した波形の数、\verb!n_af!が心房細動の出現数、\verb!group!が遺伝子導入状態（\verb!Control!と遺伝子 A を導入した\verb!A!と遺伝子 B を導入した\verb!B!の 3 水準）、\verb!id!は豚の個体 ID、\verb!day!が実験開始から何日目かを示す整数である。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/multilev2.R(1)}
\small\begin{verbatim}
# ダウンロードデータの説明
# 診断と治療社『保健医療従事者のためのマルチレベル分析活用ナビ』より
db <- read.csv("https://www.shindan.co.jp/np/filedata/00205300_13.csv")
db$group <- factor(db$group, labels=c("Control", "A", "B"))
res <- glm(n_af ~ group + day, data=db, family="poisson")
summary(res)
exp(coef(res))
exp(confint(res))
# マルチレベル
library(lmerTest)
db$afprop <- db$n_af/db$n_obs
res1 <- lmer(afprop ~ (1 | id), data=db, REML=FALSE)
summary(res1)
res2 <- lmer(afprop ~ day + group + (1 | id), data=db, REML=FALSE)
summary(res2)
res3 <- lmer(afprop ~ group + (day | id), data=db, REML=FALSE)
summary(res3)
anova(res1, res2, res3)
\end{verbatim}\end{itembox}

結果は以下。まず通常の一般化線型モデルでポアソン回帰をする\footnote{ポアソン回帰についての説明は、保健学共通特講 IV, VIII のテキストを参照されたい。}。

\begin{screen}\small\begin{verbatim}
Call:
glm(formula = n_af ~ group + day, family = "poisson", data = db)

Deviance Residuals:
    Min   1Q   Median   3Q   Max
-5.1690 -2.7101 -0.6852   1.4842   8.1568

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.96098   0.10107   9.508 < 2e-16 ***
groupA   0.69869   0.09120   7.661 1.85e-14 ***
groupB   1.31730   0.08393 15.696 < 2e-16 ***
day   0.15696   0.01479 10.612 < 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

```
    Null deviance: 1267.0  on 104  degrees of freedom
Residual deviance:  841.6  on 101  degrees of freedom
AIC: 1200.7

Number of Fisher Scoring iterations: 6

(Intercept)      groupA      groupB         day
   2.614253    2.011111    3.733333    1.169952

               2.5 %   97.5 %
(Intercept) 2.138157 3.177968
groupA      1.684723 2.409353
groupB      3.175024 4.412791
day         1.136637 1.204501
\end{verbatim}\end{screen}
```

係数が微妙に違うが、概ね藤野(2013)の図 10-3 と同じ結果が得られている。心房細動の発生頻度が、遺伝子 A の導入により 2.01 倍（95\%信頼区間 1.68-2.41）、遺伝子 B の導入により 3.73 倍（95\%信頼区間 3.17-4.42）になったことを意味する。しかし、同一個体内での心房細動発生頻度には強い相関があると考えられるので、個体の変量効果を考慮するためにマルチレベル分析をする。

\verb!lme4!パッケージは前述のように一般化線形混合効果モデルを扱えるため、ポアソン回帰のままの分析も可能なはずだが、線形混合効果モデルで分析するには反応変数を割り算して出現率にする必要がある\footnote{あまりお薦めでない方法なので、後日時間があれば、ここは一般化線形混合効果モデルで書き直す予定である。}。個体だけではなく繰り返し測定についても変量効果を考え、かつモデル間で尤度比検定をしてみた結果が以下である。

```
\begin{screen}\small\begin{verbatim}
Linear mixed model fit by maximum likelihood t-tests use Satterthwaite
  approximations to degrees of freedom [lmerMod]
Formula: afprop ~ group + (1 | id)
   Data: db

     AIC      BIC   logLik deviance df.resid
    61.1     74.4    -25.5     51.1      100

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1293 -0.6137 -0.2822  0.8290  2.6477

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.001278 0.03575
 Residual             0.094014 0.30662
Number of obs: 105, groups:  id, 15

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
```

```
(Intercept)  0.17148    0.05424 15.00000   3.162  0.00645 **
groupA       0.14824    0.07670 15.00000   1.933  0.07239 .
groupB       0.47389    0.07670 15.00000   6.178 1.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr) groupA
groupA -0.707
groupB -0.707  0.500
\end{verbatim}\end{screen}
\begin{screen}\small\begin{verbatim}
Linear mixed model fit by maximum likelihood t-tests use Satterthwaite
  approximations to degrees of freedom [lmerMod]
Formula: afprop ~ day + group + (1 | id)
   Data: db

    AIC      BIC   logLik deviance df.resid
   44.7     60.6    -16.3     32.7       99

Scaled residuals:
   Min      1Q  Median      3Q     Max
-1.9703 -0.7082 -0.1541  0.6803  3.0359

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.003761 0.06133
 Residual             0.076632 0.27682
Number of obs: 105, groups:  id, 15

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) -0.07264    0.07656 51.15000  -0.949   0.3471
day          0.06103    0.01351 90.00000   4.518 1.89e-05 ***
groupA       0.14824    0.07670 15.00000   1.933   0.0724 .
groupB       0.47389    0.07670 15.00000   6.178 1.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr) day    groupA
day    -0.706
groupA -0.501  0.000
groupB -0.501  0.000  0.500
\end{verbatim}\end{screen}
\begin{screen}\scriptsize\begin{verbatim}
Linear mixed model fit by maximum likelihood t-tests use Satterthwaite
  approximations to degrees of freedom [lmerMod]
Formula: afprop ~ group + (day | id)
```

```
  Data: db

    AIC      BIC   logLik deviance df.resid
   31.2     49.7     -8.6     17.2       98

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.11685 -0.61717 -0.07123  0.44926  2.60950

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 0.18725  0.4327
          day         0.01018  0.1009   -0.98
 Residual             0.04651  0.2157
Number of obs: 105, groups:  id, 15

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  0.16353    0.05414 15.00000   3.021   0.0086 **
groupA       0.17168    0.07656 15.00000   2.242   0.0405 *
groupB       0.50530    0.07656 15.00000   6.600 8.44e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr) groupA
groupA -0.707
groupB -0.707  0.500
\end{verbatim}\end{screen}
\begin{screen}\small\begin{verbatim}
Data: db
Models:
object: afprop ~ group + (1 | id)
..1: afprop ~ day + group + (1 | id)
..2: afprop ~ group + (day | id)
       Df    AIC    BIC   logLik deviance  Chisq Chi Df Pr(>Chisq)
object  5 61.088 74.358 -25.5439   51.088
..1     6 44.689 60.613 -16.3444   32.689 18.399      1  1.792e-05 ***
..2     7 31.151 49.729  -8.5756   17.151 15.538      1  8.088e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
\end{verbatim}\end{screen}
```

尤度比検定の結果、グループの固定効果と個体の変量効果のみ考えた最初のモデルに比べて、日の固定効果も含めた２番目のモデルや、日の効果に個体差を考慮した最後のランダム切片モデル（\verb!Cor!が\verb!-0.98!と個体差が日数と強い負の相関があることもわかる）は、どちらも 5\%水準で（p 値はどちらも10^{-5}のオーダーだから 5\%どころではないが）有意に当てはまりが良く、\verb!AIC!の値を見ると最後のモデルが最も小さい値を示しているので、最も当てはまりが良いと考えられる。

Example 3: Support in different workplaces

藤野(2013)pp.75-76 に掲載されている、企業における血圧と職場サポートの関連の検討である。職場サポートは質問票により部署ごとに計算された平均点をその職場のサポートスコアとした。データに含まれている変数は、\verb!workplace!が部署、\verb!bp_s!が収縮期血圧、\verb!age!が年齢、\verb!support!がサポートスコアである。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/multilev2.R(2)}
\small\begin{verbatim}
wps <- read.csv("https://www.shindan.co.jp/np/filedata/00205300_16.csv")
wps$workplace <- as.factor(wps$workplace)
res1 <- lm(bp_s ~ age + support, data=wps)
summary(res1)
PCHS <- c(1:9, LETTERS)
plot(bp_s ~ age, pch=PCHS[as.integer(workplace)], data=wps,
 col=ifelse(support<median(support), "red", "green"))
library(lmerTest)
res4 <- lmer(bp_s ~ age + support + (1 | support), data=wps, REML=FALSE)
summary(res4)
confint(res4)
res5 <- lmer(bp_s ~ age + (1 | support), data=wps, REML=FALSE)
summary(res5)
confint(res5)
anova(res4, res5)

print(devdif <- as.numeric(-2*(logLik(res1)-logLik(res4))))
print(dfdif <- attr(logLik(res4), "df")-attr(logLik(res1), "df"))
pchisq(devdif, dfdif, lower.tail=FALSE)
\end{verbatim}\end{itembox}

まずはサポートの有無と年齢の血圧への効果をみる線形回帰モデルの結果を示す。

\begin{screen}\small\begin{verbatim}
Call:
lm(formula = bp_s ~ age + support, data = wps)

Residuals:
    Min   1Q Median   3Q   Max
-40.603 -10.943 -1.209   9.676 97.977

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.02510   9.67731 13.023 < 2e-16 ***
age   0.42023   0.03472 12.102 < 2e-16 ***
support   -3.64080   1.17204 -3.106 0.00193 **

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.31 on 1453 degrees of freedom
Multiple R-squared:  0.108,     Adjusted R-squared:  0.1068
F-statistic: 87.96 on 2 and 1453 DF,  p-value: < 2.2e-16
\end{verbatim}\end{screen}
```

$$\includegraphics[width=10cm]{wps-support.pdf}$$

\verb!support!の係数が-3.6、\verb!age!の係数が 0.42 でともに有意水準 5\%で有意だが、モデル全体の説明力は低く、自由度調整済重相関係数の二乗（\verb!Adjusted R-squared!）が 10.7\%しかない。

この場合、血圧に影響する要因は部署ごとに異なっているだろうし、個人差もあるだろうと考える方が自然である。それらの変量効果を考えるマルチレベルモデルの結果を以下示す。職場単位でネストされた構造（レベル 1 が各対象者個人、レベル 2 が部署）とする。

```
\begin{screen}\scriptsize\begin{verbatim}
Linear mixed model fit by maximum likelihood t-tests use Satterthwaite approximations to
  degrees of freedom [lmerMod]
Formula: bp_s ~ age + support + (1 | support)
   Data: wps

     AIC      BIC   logLik deviance df.resid
 12040.6  12067.0  -6015.3  12030.6     1451

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9741 -0.6951 -0.0728  0.6009  6.3136

Random effects:
 Groups   Name        Variance Std.Dev.
 support  (Intercept)   9.666   3.109
 Residual             222.703  14.923
Number of obs: 1456, groups:  support, 34

Fixed effects:
             Estimate Std. Error         df t value Pr(>|t|)
(Intercept)  130.42946   16.23625   39.70000   8.033 7.54e-10 ***
age            0.39737    0.03715 1202.90000  10.697  < 2e-16 ***
support       -4.08753    1.99687   37.00000  -2.047   0.0478 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
        (Intr) age
age     -0.224
support -0.994  0.124
```

```
Computing profile confidence intervals ...
                2.5 %      97.5 %
.sig01       2.0576492   4.5851929
.sigma      14.3931343  15.4866022
(Intercept) 98.3737422 163.9122155
age          0.3244992   0.4702249
support     -8.2163934  -0.1418429
\end{verbatim}\end{screen}
```

藤野ら(2013)の図 10-6 とほぼ同じ結果が得られた。以上の結果は下表のようにまとめることができる。

```
$$\vbox{
\begin{tabular}{lrrrr}
\noalign{\noindent 表. 収縮期血圧への年齢と職場サポートの固定効果と職場サポートの変量効果を考慮したマルチレベル分析}
\hline
固定効果の変数 & 係数 & 標準誤差 & 95\%下限 & 95\%上限 & p 値\cr
\hline
年齢(\verb!age!) & 0.397 & 0.037 & 0.324 & 0.470 & $<$0.001 \cr
サポート(\verb!support!) & $-0.409$ & 1.997 & $-8.216$ & $-0.142$ & 0.048\cr
切片 & 130.4 & 16.2 & 98.4 & 163.9 & $<$0.001\cr
\hline
固定効果間の相関\cr
\hline
年齢-切片 & $-0.224$\cr
サポート-切片 & $-0.994$\cr
サポート-年齢 & 0.124\cr
\hline
変量効果の変数 & 標準偏差 & 95\%下限 & 95\%上限\cr
\hline
サポート & 3.11 & 2.06 & 4.59\cr
個人差 & 14.92 & 14.39 & 15.49\cr
\hline
\noalign{\noindent AIC=12040.6, N=1456, 部署数=34}
\end{tabular}
}$$
```

最後に\verb!support!の変量効果だけを考え、固定効果を考えないモデルと尤度比検定すると、以下のように\verb!support!の固定効果を入れたモデルの方が有意に当てはまりが良いといえた。

```
\begin{screen}\scriptsize\begin{verbatim}
Linear mixed model fit by maximum likelihood t-tests use Satterthwaite approximations to
  degrees of freedom [lmerMod]
Formula: bp_s ~ age + (1 | support)
   Data: wps

     AIC      BIC   logLik deviance df.resid
```

```
 12042.7  12063.8  -6017.3  12034.7     1452

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9151 -0.6928 -0.0702  0.5965  6.3649

Random effects:
 Groups   Name        Variance Std.Dev.
 support  (Intercept)  10.8     3.286
 Residual             223.0    14.935
Number of obs: 1456, groups:  support, 34

Fixed effects:
            Estimate Std. Error        df t value Pr(>|t|)
(Intercept) 9.743e+01  1.833e+00 3.200e+02   53.14   <2e-16 ***
age         4.063e-01  3.701e-02 1.121e+03   10.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
    (Intr)
age -0.903

Computing profile confidence intervals ...
                2.5 %      97.5 %
.sig01       2.1696680   4.8594543
.sigma      14.4037990  15.4990512
(Intercept) 93.8309194 101.0484448
age          0.3335201   0.4789125

Data: wps
Models:
..1: bp_s ~ age + (1 | support)
object: bp_s ~ age + support + (1 | support)
       Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
..1     4 12043 12064 -6017.3    12035
object  5 12041 12067 -6015.3    12031 4.1136      1    0.04254 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
\end{verbatim}\end{screen}
```

藤野ら(2013)では、\verb!res4!に得られている年齢とサポートの固定効果とサポートの変量効果を考えた混合効果モデルを、\verb!res1!に得られている年齢とサポートから血圧への線形回帰モデルと尤度比検定で比べている。ただし、\verb!lm()!と\verb!lmer()!の結果オブジェクトのクラスが違うので、R では\verb!anova()!関数に渡すだけで尤度比検定を実行することはできない。
To solve this issue, manually calculating the likelihood ratio and degree of freedom, then applying chi-square distribution[31] as shown below.

31 https://www.cl.uni-heidelberg.de/statnlpgroup/empirical_methods_tutorial/icml2022_slides/LMEM.pdf

\begin{screen}\small\begin{verbatim}
> print(devdif <- as.numeric(-2*(logLik(res1)-logLik(res4))))
[1] 43.88222
> print(dfdif <- attr(logLik(res4), "df")-attr(logLik(res1), "df"))
[1] 1
> pchisq(devdif, dfdif, lower.tail=FALSE)
[1] 3.487428e-11
\end{verbatim}\end{screen}

Example 4: Built-in data in R

\verb!lmer!パッケージの\verb!sleepstudy!というデータを使う。180 オブザーベーション、3 変数からなり、含まれている変数は\verb!Reaction!、\verb!Days!、\verb!Subject!である。このデータは、健康なボランティアを対象にして、睡眠時間を奪うと反応時間がだんだん長くなっていくことを検証したものである（Belenky {\it et al.\/}, 2003)\footnote{\url{https://doi.org/10.1046/j.1365-2869.2003.00337.x}からフルテキスト読める。}。実験 0 日目には普通に睡眠をとってもらい、翌日から 3 時間に睡眠時間を制限する（元論文ではベースラインは 3 日間 8 時間睡眠、実験期間中は 3 時間の他に、5 時間、7 時間、9 時間という実験条件を 4 群でそれぞれ 7 日間続け、最後に 3 日間 8 時間睡眠としているが、このデータは 3 時間睡眠実験群しか含んでいない）。反応時間は、LED を使った視覚刺激提示後に指で反応するまでの時間をミリ秒単位で計測している。

\begin{itembox}[l]{sleepstudy.R}\small\begin{verbatim}
if (require(lme4)==FALSE) {
 install.packages("lme4", dep=TRUE); library(lme4) }
data(sleepstudy)
str(sleepstudy)
bysubjects <- split(sleepstudy[, 2:1], sleepstudy[, 3])
plot(bysubjects[[1]], type="b", ylim=c(150, 500))
for (i in 2:length(bysubjects)) points(bysubjects[[i]], type="b", pch=i)
res1 <- lmer(Reaction ~ Days + (Days | Subject), data=sleepstudy)
summary(res1)
\end{verbatim}\end{itembox}

$$\includegraphics[width=10cm]{sleepstudy.pdf}$$

\begin{screen}\small\begin{verbatim}
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy

REML criterion at convergence: 1743.6

Scaled residuals:
    Min   1Q Median   3Q   Max
-3.9536 -0.4634 0.0231 0.4634 5.1793

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.09   24.740
          Days         35.07    5.922   0.07
 Residual             654.94   25.592
Number of obs: 180, groups:  Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84
Days          10.467      1.546    6.77

Correlation of Fixed Effects:
     (Intr)
Days -0.138
\end{verbatim}\end{screen}
```

ランダム効果の標準偏差の推定値は、切片と傾きについてそれぞれ一日当たり 24.74 ミリ秒と 5.92 ミリ秒である。固定効果の係数は、切片と傾きについて、それぞれ一日当たり 251.4 ミリ秒と 10.47 ミリ秒である。

Controlling many confounding factors

When we have to control many confounding factors simultaneously, some sophisticated approaches have been proposed recently: Propensity Score Matching (PSM), Difference in Differences (DID), and Instrumental Variables regression. Most of those have been developed in econometric studies.

PSM

The concept of the propensity score was proposed by Rosenbaum and Rubin in 1983[32]. Before explaining propensity scores, let's first discuss the Average Treatment Effect as a causal effect.
In observational studies, random assignment is impossible. However, it's not impossible to estimate the Average Treatment Effect (ATE), which is the expected difference in outcomes in a population between cases where a certain factor or treatment is present and when it's absent. Of course, in reality, a population contains a mix of individuals who have and do not have the factor or treatment. For someone who has the factor or treatment, the "case where it wasn't present" is a counterfactual. Therefore, direct estimation is not possible unless random assignment can be performed.
However, if we perform matching, treating person j as the same as person i when person j has covariate values identical or similar to person i (who has the factor or treatment), and then create pairs of treated and control groups, we can estimate the causal effect as the difference in outcomes between these two groups, similar to random assignment. The problem is that if there are many covariates (confounding factors), it becomes practically impossible to find matches where all those covariates are identical or very close in value. This is where the propensity score comes in. In the example above, the probability that person i is in the treatment group is called person i's propensity score. Since it's a probability, the

32 This original article is available at https://doi.org/10.2307/2335942. I recommend you to read the newer introductory article such as Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, *46*(3), 399-424. https://doi.org/10.1080/00273171.2011.568786.

propensity score is a one-dimensional variable taking values in [0, 1]. If the "strong ignorability" assumption (where group assignment and potential outcomes are independent given the covariates) is satisfied, adjustment using propensity scores becomes possible. If we match individuals based on their calculated propensity scores, we can correctly estimate the Average Treatment Effect. Typically, predicted probabilities obtained from logistic regression or probit regression analysis can be used as propensity scores.
There are various ways to utilize the calculated propensity scores for each individual. These include matching or stratification based on the score, using the propensity score as an explanatory variable in kernel regression models, inverse probability weighting methods using propensity scores, and "doubly robust estimation methods," among others.

DID

Recently, many introductory papers are published for DID, including Callaway and Sant'Anna (2021)[33].Here, I will explain the concept of DID based on the explanation by Yamaguchi (2016).

育児支援が女性の労働に及ぼす効果の研究において、一般に認可保育所の整備が女性の就業率を上げると思われているが、ノルウェー、フランス、米国等での先行研究によると、公的保育の整備にもかかわらず母親の就業は増えなかったという報告があるので、認可保育所の整備が母親就業につながるかを日本のデータで検証したという話である。彼らは認可保育所整備と女性就業率の都道府県間格差に注目した。まず、0-5 歳の子供 1 人当たりの認可保育所定員数を保育所定員率と名付け、これを横軸にとって、0-5 歳の子供をもつ母親の就業率を縦軸にとってプロットすると正の相関関係が見られることを示し、このことが認可保育所の整備が母親就業率を上げるという印象を与えることに触れた後で、それが県民性の違いによる（母親の就業意欲が高く地域社会がそれに対して好意的ならば母親就業率も上がるし政治的支持を得やすいので保育所整備も進む）という可能性を指摘し、その解析のために縦軸横軸とも 2005 年から 2010 年まで 5 年間の変化（階差）をとってプロットすることによって県民性の影響を排除すると（もし認可保育所の整備によって母親就業率が上がるならば変化同士も相関しているはずなのに）相関が消えてしまうことを示した。

そこで DID を使うには、状況を単純化する。どの都道府県でも保育所定員率が増えているけれども、大きく増えた都道府県と少ししか増えなかった都道府県に二分して考え、前者が保育所を増やすという処置をした結果であると考えて、処置群と呼ぶことにする（残りが対照群）。処置群、対照群について、それぞれ各時点における 0-5 歳の子供をもつ家計数で重み付けした平均母親就業率を求め、2005 年の処置群、2010 年の処置群、2005 年の対照群、2010 年の対照群の順に、その値を A、B、C、D と書くと、2010 年における保育所定員率と母親就業率に有意な相関があることは、B と D に有意差があることに相当する。対照群は無視して処置前後で母親就業率に差があるかをみるという A と B の比較では、保育所整備以外の経済・社会的情勢の変化の効果も含まれてしまうので、その部分、つまり経済・社会的情勢の変化の効果が、処置がなかったところでの母親就業率の変化である C と D の差に現れると考えれば、もし経済・社会的情勢が変化しなかったら処置の効果はどうなるか、つまり$(B-A)-(D-C)$によって処置の効果を評価することができる。これが「差の差」である。

実際に処置効果を推定するための回帰モデルは、都道府県pにおけるt年の母親就業率をY_{pt}とし、D^T_pを都道府県pが処置群に属するかどうかを示すダミー変数（属していれば 1、対照群なら 0）とし、D_{t+1}を年次ダミー（t年のデータに対して 0、$t+1$年のデータに対して 1）として、

33 Callaway B, Sant'Anna PHC (2021) Difference-in-Differences with multiple time periods. *Journal of Econometrics*, 225(2): 200-230. https://doi.org/10.1016/j.jeconom.2020.12.001.

$$Y_{pt} = \alpha + \beta D^T_p D_{t+1} + \gamma D^T_p + \delta D_{T+1} + \epsilon_{pt}$$
を推定すれば良いとのことなので（{\bf 処置効果はβが有意かどうか、経済・社会情勢の変化の効果はδが有意かどうかで見ることができる}）、\verb!lm()!で分析可能なはずである。

なお、就業構造基本調査で「育児をしている」という区分は未就学児についての集計であり、平成 24 年の未就学児の育児をしている女性のデータは\url{https://www.e-stat.go.jp/SG1/estat/GL08020103.do?_xlsDownload_&fileId=000006464031&releaseCount=1}から Excel 形式で得られるが、ここで使われているのは国勢調査と書かれていて、2005 年と 2010 年のデータとのことなので、e-Stat で探してみたが、どこにあるのかわからなかったため\footnote{e-Stat から入手できると書くだけではなく、データベース名あるいは URL を明記しておいて欲しいところ。}、実際の分析を示すことはできない。

{\it Annual Reviews of Public Health\/}に、Wing C et al. "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research"\footnote{\url{https://www.annualreviews.org/doi/10.1146/annurev-publhealth-040617-013507}}という論文が 2018 年に掲載されていて、公衆衛生政策の評価についても DID 解析を使う研究デザインが役に立つことが示されている。

R での DID の実行方法については、プリンストン大学のサイトに演習用プレゼンテーションファイルが掲載されている[34]。これはグループと時点の交互作用を示す変数を作って、線形回帰モデル\verb!lm()!を使う方法なので、とくにパッケージなどは必要としない。A から G の 7 つの国について 1990 年から 1999 年までの 10 年間の何かの量（\verb!y!）が示されていて、E、F、G の 3 つの国では何らの処置がとられ、処置は 1994 年からとられたと想定し、処置がとられた国ととられていない国で\verb!y!の変化に違いがあるかを DID 法で評価している、コードは以下の通り。

\begin{itembox}[l]{}\begin{verbatim}
library(foreign)
dat <- read.dta("https://dss.princeton.edu/training/Panel101.dta")
dat$time <- ifelse(dat$year>=1994, 1, 0)
dat$treated <- ifelse(dat$country %in% LETTERS[5:7], 1, 0)
dat$did <- dat$time * dat$treated
didreg <- lm(y ~ treated + time + did, data=dat)
summary(didreg)
\end{verbatim}\end{itembox}

結果は以下の通り。

\begin{screen}\small\begin{verbatim}
Call:
lm(formula = y ~ treated + time + did, data = dat)

Residuals:
      Min   1Q   Median   3Q   Max
-9.768e+09 -1.623e+09 1.167e+08 1.393e+09 6.807e+09

Coefficients:

34 https://www.princeton.edu/~otorres/DID101R.pdf.
The R code is available at https://minato.sip21c.org/advanced-statistics/princetondid.R.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.581e+08  7.382e+08   0.485   0.6292
treated      1.776e+09  1.128e+09   1.575   0.1200
time         2.289e+09  9.530e+08   2.402   0.0191 *
did         -2.520e+09  1.456e+09  -1.731   0.0882 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.953e+09 on 66 degrees of freedom
Multiple R-squared:  0.08273,   Adjusted R-squared:  0.04104
F-statistic: 1.984 on 3 and 66 DF,  p-value: 0.1249
\end{verbatim}\end{screen}
```

に示す通り、\verb!time!の係数が正で p 値が 0.0191 と 5\%有意なので、時間経過によって\verb!y!は増加していると言えるが、\verb!did!の p 値が 0.088 と 0.05 より大きいので、時点と処置の交互作用項である\verb!did!が有意に\verb!y!に影響しているとは言えない。つまり、処置があってもなくても、\verb!y!が経時的に増加したことに影響はなかったと考えられる。

R で DID を実行するために開発されたパッケージとして\verb!did!\footnote{\url{https://cran.r-project.org/web/packages/did/did.pdf}}があり、cran からインストールできるし、開発者による使い方の解説記事も発表されている\footnote{\url{https://bcallaway11.github.io/did/articles/did-basics.html}}。

その他の情報としては、\url{https://thetarzan.wordpress.com/2011/06/20/differences-in-differences-estimation-in-r-and-stata/}や、\url{https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation}や\url{https://static1.squarespace.com/static/59371c8ad1758ebe90723e40/t/5cdf25f58f78a100018f804f/1558128117692/strumpf+2017-DD+and+FE.pdf}も参考になる。

Instrumental Variables regression

操作変数法は、構造方程式モデルのところで紹介した Dr. John Fox のチュートルアル文書の中で\verb!sem!パッケージの\verb!tsls()!関数を使った方法が説明されているが、\verb!AER!パッケージの\verb!ivreg()!関数を推奨する。

\subsection{二段階最小二乗法の関数 tsls()による操作変数法}

組み込みデータ\verb!Klein!を用いる。このデータは 1921 年から 1941 年の米国経済について Klein が発表した単純な経済測定モデルに使われている。変数の意味は以下の通りである。

\begin{screen}\small\begin{description}
\item[Year] 1921-1941
\item[C] consumption.
\item[P] private profits.
\item[Wp] private wages.
\item[I] investment.
\item[K.lag] capital stock, lagged one year.
\item[X] equilibrium demand.

\item[Wg] government wages.
\item[G] government non-wage spending.
\item[T] indirect business taxes and net exports.
\end{description}\end{screen}

\begin{itembox}[l]{\url{https://minato.sip21c.org/advanced-statistics/tsls.R}}\small\begin{verbatim}
library(sem)
data(Klein)
Klein$P.lag <- c(NA,Klein$P[-22])
Klein$X.lag <- c(NA,Klein$X[-22])
# model 1
Klein.eqn1 <- tsls(C ~ P + P.lag + I(Wp+Wg),
 instruments=~G + T + Wg + I(Year-1931) + K.lag + P.lag + X.lag, data=Klein)
summary(Klein.eqn1)
plot(Klein$Year[-1], residuals(Klein.eqn1))
# model 2
Klein.eqn2 <- tsls(I ~ P + P.lag + K.lag,
 instruments=~G + T + Wg + I(Year-1931) + K.lag + P.lag + X.lag, data=Klein)
summary(Klein.eqn2)
plot(Klein$Year[-1], residuals(Klein.eqn2))
# model 3
Klein.eqn3 <- tsls(Wp ~ X + X.lag + I(Year-1931),
 instruments=~G + T + Wg + I(Year-1931) + K.lag + P.lag + X.lag, data=Klein)
summary(Klein.eqn3)
plot(Klein$Year[-1], residuals(Klein.eqn3))
\end{verbatim}\end{itembox}

How to use ivreg function of AER package

\verb!AER!パッケージの\verb!CigarettesSW!というデータフレームは、米国の州ごとのタバコ消費量と価格や税金等の関連因子のデータを含んでいる。詳細は\verb!?CigarettesSW!とプロンプトに打てば表示される。

\begin{itembox}[l]{https://minato.sip21c.org/advanced-statistics/ivreg.R}\small\begin{verbatim}
library(AER)
data("CigarettesSW")
CigarettesSW$rprice <- with(CigarettesSW, price/cpi)
CigarettesSW$rincome <- with(CigarettesSW, income/population/cpi)
CigarettesSW$tdiff <- with(CigarettesSW, (taxs-tax)/cpi)
CigarettesSW$rtax <- with(CigarettesSW, tax/cpi)
CigarettesSW$lrprice <- log(CigarettesSW$rprice)
CigarettesSW$lrincome <- log(CigarettesSW$rincome)
CigarettesSW$lpacks <- log(CigarettesSW$packs)
CSW1995 <- subset(CigarettesSW, year=="1995")

fm <- ivreg(lpacks ~ lrprice + lrincome | lrincome + tdiff + rtax, data=CSW1995)
summary(fm)

```
fm2 <- ivreg(lpacks ~ lrprice | tdiff, data=CSW1995)
anova(fm, fm2)

library(sem)
M1 <- tsls(lpacks ~ lrprice + lrincome,
 instruments = ~ lrincome + tdiff + rtax, data=CSW1995)
summary(M1)
\end{verbatim}\end{itembox}
```

上記コードを実行すると、まず、\verb!summary(fm)!で以下が表示される。

```
\begin{screen}\small\begin{verbatim}
Call:
ivreg(formula = lpacks ~ lrprice + lrincome | lrincome + tdiff +
    rtax, data = CSW1995)

Residuals:
      Min        1Q    Median        3Q       Max
-0.6006931 -0.0862222 -0.0009999  0.1164699  0.3734227

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8950     1.0586   9.348 4.12e-12 ***
lrprice      -1.2774     0.2632  -4.853 1.50e-05 ***
lrincome      0.2804     0.2386   1.175    0.246
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1879 on 45 degrees of freedom
Multiple R-Squared: 0.4294,  Adjusted R-squared: 0.4041
Wald test: 13.28 on 2 and 45 DF,  p-value: 2.931e-05
\end{verbatim}\end{screen}
```

次に、\verb!anova(fm, fm2)!で以下が表示される。

```
\begin{screen}\small\begin{verbatim}
Analysis of Variance Table

Model 1: lpacks ~ lrprice + lrincome | lrincome + tdiff + rtax
Model 2: lpacks ~ lrprice | tdiff
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     45 1.5880
2     46 1.6668 -1 -0.078748 1.3815  0.246
\end{verbatim}\end{screen}
```

最後に\verb!summary(M1)!で以下が表示される。このように\verb!sem!パッケージの\verb!tsls()!関数を使っても、\verb!AER!パッケージの\verb!ivreg()!関数を使った場合に得られる\verb!summary(fm)!の出

力とほぼ同じ結果が得られるが、自由度調整済重相関係数の二乗やウォルドの検定結果は表示されないので、\verb!AER!パッケージの使用を推奨する。

\begin{screen}\small\begin{verbatim}
 2SLS Estimates

Model Formula: lpacks ~ lrprice + lrincome

Instruments: ~lrincome + tdiff + rtax

Residuals:
      Min.   1st Qu.   Median   Mean   3rd Qu.   Max.
-0.6006931 -0.0862222 -0.0009999 0.0000000 0.1164699 0.3734227

             Estimate Std. Error t value   Pr(>|t|)
(Intercept) 9.8949555 1.0585599 9.34756 4.1209e-12 ***
lrprice   -1.2774241 0.2631986 -4.85346 1.4960e-05 ***
lrincome   0.2804048 0.2385654 1.17538   0.24602

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.187856 on 45 degrees of freedom
\end{verbatim}\end{screen}

References

About R

Overview

R-project: https://www.r-project.org/
Download (CRAN): https://cran.r-project.org/
Package search (METACRAN): https://www.r-pkg.org/
News (R-bloggers): https://www.r-bloggers.com/

About statistics

Upton G, Cook I (2014) A Dictionary of Statistics, 3rd Ed. Oxford University Press, ISBN: 9780199679188.

Factor analysis

Everitt B, Hothorn T (2011) An introduction to applied multivariate analysis in R. Springer. Full text at https://link.springer.com/book/10.1007/978-1-4419-9650-3.

SEM

E-learning course at UCLA: https://stats.oarc.ucla.edu/r/seminars/rsem/
Jongering J et al. (2024) Structural equation modeling with R for educational scientists. In “Learning analytics methods and tutorials” Springer. Full text at https://link.springer.com/chapter/10.1007/978-3-031-54464-4_21.

Multilevel models

Faraway JJ (2006) Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models, Chapman and Hall (especially Chapter 8).

Propensity score and instrumental variable

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika, 70: 41-55.