Medical Statistics for Gunma Univ. Graduate School of Medicine

Minato Nakazawa, Ph.D. (Dept. Public Health, Assoc. Prof.)

May 25 and June 1, 2011

The purpose of this practice is to master the following series of data analysis: (1) to make computerized data file from raw data collected by experiments or field survey, (2) to analyze the data using the free software **R**, (3) to read the results and (4) to summarize them as a report.

Contents

Corresponding to: Minato Nakazawa, Ph.D., Assoc. Prof. of Dept. Public Health, Gunma Univ.
e-mail: nminato@med.gunma-u.ac.jp

Rev. 0.5 on 26 May 2010: Until half.
Rev. 1.0 on 17 August 2010: Completed. (However, English brush-up is needed!)
Rev. 1.1.1 on 10 November 2010: Revised and the survival analysis was added (except Cox regression).
Rev. 1.1.2 on 16 November 2010: Cox regression was added.
Rev. 1.1.3 on 22 November 2010: Wrong translation into Japanese at the exercise of logistic regression was corrected.
Rev. 1.1.4 on 20 May 2011: Description of R version and dates were updated and explanation about multiple imputation relating missing values was added.
Rev. 1.1.4.1 on 25 May 2011: Minor correction.
Rev. 1.1.4.2 on 1 June 2011: Minor correction.

1 The very basics of R

R can work on many computer operating systems like Microsoft Windows, Mac OS X, or Linux. To install R on Windows and Mac OS X, we can download the appropriate binary setup file and execute it with selecting some options. To install R on Linux, we can usually download source tar ball and make and install.

R is a free software, so that you can freely install and use it in your own computer. The internet sites where we can download R-related files (including binary setup files and source tar balls, with many additional packages) are called as "CRAN" (The Comprehensive R Archive Network). There are 2 mirror sites of CRAN in Japan and residents in Japan are recommended to use them (Univ. Tsukuba [*1] and Hyogo University of Teacher Education[*2]).

1.1 Installation of R programs, as of 20 May 2011

Windows Download `R-2.13.0-win.exe` from CRAN mirror and execute it. English is recommended as language used in the process of installation. In the window "Setup - R for Windows 2.13.0", click [Next], then you see the license confirmation. Click [Next] again, you must specify the directory for R programs. In general, the default, `C:\Program Files\R\R-2.13.0`, is recommended. Clicking [Next], you must select the components to be installed. Again, in general, the default "User installation" is enough (But the author of this text uses "Full installation". Then clicking [Next], the window to ask "Do you want to customize the startup options?" appears. Here, you should select **Yes (customized startup)**, because accepting defaults force you to use MDI environments which will make conflicts with Rcmdr. In the next window, **SDI (separate windows)** should be checked. Clicking [Next], you must select the type of help system. Either OK, but the author prefer "Plain text". In the next window, you must specify the internet connection type. If you have already set up Internet Explorer to access any internet sites, "Internet2" is recommended. After that, explanation may not be needed.

Macintosh R-2.13.0 can work on Mac OS X 10.5 (Leopard) or later versions. Downloading `R-2.13.0.pkg` from R mirror sites and double-click it.

Linux You can download pre-compiled binaries for major distribution packages like Debian, Redhat or ubuntu from CRAN. Otherwise please download `R-2.13.0.tar.gz` from CRAN mirror sites and execute `tar xvzf R-2.13.0.tar.gz`. Changing directory to R-2.13.0 and doing `./configure` and `make`, you can get executable binary. After that, you must become superuser before doing `make install`.

1.2 Basic usage of R

The following description is based on Windows environment. There may be some environment-specific points.

You will find "R" icon on your desktop after the completion of installation. If you cannot find it on your Desktop, you can find it in the start menu. To start R gui (graphical user's interface), just double-click this icon, then R gui environment will start. If you previously set the working directory of this icon's property to your working directory, R uses there as the current directory. **And, if you set "Link to" box as** `["C:\Program Files\R\R-2.13.0\bin\i386\Rgui.exe" LANG=C LC_ALL=C]`, **you can use R in English mode even in Japanese Windows.** The Rgui executes `.Rprofile` of the working directory and reads `.RData`, and shows the following prompt.

```
>
```

Entering commands (functions) to R should be done via this prompt. When you press the Enter key on the middle of line, the prompt will change to +. In the Windows environment, you can cancel the command to back to the first prompt >

[*1] `http://cran.md.tsukuba.ac.jp/`

[*2] `http://essrc.hyogo-u.ac.jp/cran/`

by pressing Esc key.

You can save the history which you entered as functions, statements, comments (R will treat any statements as comments after # until the end of line) into a file (default name is `.Rhistory`. You can redo any saved functions by entering `source("saved filename")`. The directory delimiter sign should be / instead of \[*3].

You can recall any lines which you have entered in that session by pressing ↑ key.

If you add `C:\Program Files\R\R-2.13.0\bin` to Path, you can start R by simply typing R in the command prompt of Windows 2000/XP.

1.3 Basic functions to be entered to the Rgui prompt

Quit `q()`

Assign `<-`

For example, to assign the numeric vector with 3 elements of 1, 4, 6 to the variable X, type as follows.

```
X <- c(1,4,6)
```

Define function `function()`

For example, the combined function of `mean()` and `sd()`, `meansd()` can be defined as follows.

```
meansd <- function(X) { list(mean(X),sd(X)) }
```

Install packages `install.packages()`

For example, downloading the Rcmdr package with depending-on packages from CRAN can be done as follows[*4],

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

Help `?`

For example, to see the help of `t.test` (which statistically tests the null-hypothesis that there is no significant difference between the means of two independent groups), type `?t.test` to the prompt.

The function definition has great possibility. It can be done using many lines, and the return value of the function is the last line of the definition, and the return value can be any object: not only scalar, but also vector, matrix, list, or data.frame.

Each function has its own scope. Assignment within a function has no effect outside, unless using eternal assignment by `<<-`.

1.4 Using R Commander

This practical course has so limited time to learn R that such a command-based usage is unsuitable. Therefore, we use the `Rcmdr` package, which enhance the menu-based graphical users' interface. To start `Rcmdr`, type `library(Rcmdr)` to the Rgui's prompt. After you once terminate Rcmdr, you need to type `detach(package:Rcmdr)` before restart the Rcmdr by typing `library(Rcmdr)` again. However, it is also possible to call Rcmdr by typing `Commander()` without detaching the package.

[*3] This character is ¥ in Japanese keyboard.

[*4] However, to install additional packages to R in Windows environments, the Administrator's right is necessary.

2 Data entry, descriptive statistics, and drawing graph

2.1 Data entry

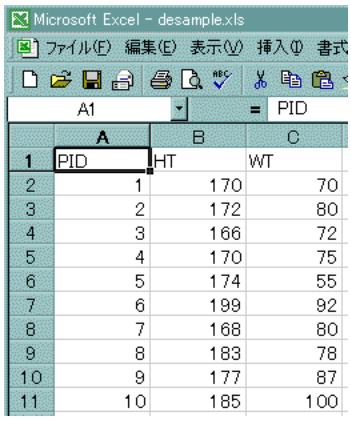
For the statistical analysis of the data obtained through research, at first you must enter the data into the computer. The suitable (accurate and efficient) way to enter the data depends on the size of the data and the software to statistically analyze data.

If your data is very very limited and analysis is very very simple, you can use even calculator, not computer. At least, you don't have to make data file, just type the data values within a procedure of analysis. For example, mean body weight of 3 individuals of 60, 66, 75 kilograms can be calculated by typing `mean(c(60,66,75))` or `(60+66+75)/3` to R's prompt.

However, most researchs require much bigger sized data analysis with various method. In such cases, we should prepare the data file, separated from analyzing program. Somebody uses the Microsoft Excel for both data entry and analysis, both can be entered into cells in similar manner, but I don't recommend it from the view of protection and secure management and future re-analysis possibility.

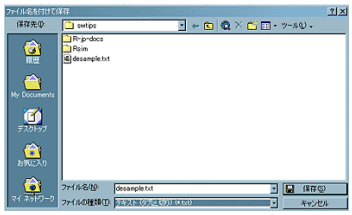
Spreadsheet programs like Microsoft Excel or OpenOffice.org Calc should be exclusively used for data entry. For example, the following table is the data of weight and height for 10 subjects.

Subject ID	Height (cm)	Weight (kg)
1	170	70
2	172	80
3	166	72
4	170	75
5	174	55
6	199	92
7	168	80
8	183	78
9	177	87
10	185	100



In a spreadsheet software, this table should be entered into a single sheet. The top raw should be variable names. Multi-byte characters can be used as variable name, but ASCII characters (especially alphabets and period, case-sensitive) can be recommended. Actually some special characters are not allowed as R object name; for example, _ (underscore), #, and so on. If these special characters are included in the top raw of the file, R will automatically change them, but it may make trouble. After completing data entry, the file should be once saved as the software's standard format (*.xls in Excel, *.ods in Calc). The screenshot shown in the right is an example.

Next, we must save this as tab-delimited text file. From "File (F)" menu, select "Save as" and specify file format as text file (delimited by tab): For example, name it as `desample.txt`. Though some warning dialog boxes will appear, you can ignore them and click [Yes] button. The text file should not be placed on Desktop in Japanese environment, because Rcmdr in English version sometimes fails to read file on the directory with the name including Japanese characters (but "My Documents" is OK).



Next, we will read `desample.txt` into R. In Rgui console, we can simply type as follows, then the data in the tab-delimited text file will be imported into the R's data.frame object `Dataset`. The data.frame object includes many named variables of same length. You can use any possible object name as the name of data.frame.

```
Dataset <- read.delim("desample.txt")
```

In Rcmdr, select [Data] in menu bar, [Import data], and [from text file, clipboard, or URL ...]. Then enter an arbitrary name (default, "Dataset") into the text box of [Enter name for data set:], check [Tabs] radio button as [Field Separator] and click [OK] button. After that, you will see the window to select data file. You can confirm the successful reading by clicking [View data set] button.
The spreadsheet's data can also be read by Rcmdr via clipboard, without making any file. Just after the completion of data entry, you select the all data ranges and copy them to the clipboard. Activating Rcmdr window and after selecting [Data], [Import data], [from text file, clipboard, or URL, ...] and entering appropriate name into [Enter name for data set:] and check [Tabs] radio button as [Field Separator], check [Clipboard] radio button, and click [OK] button, then the data will be set as active [Data set] in Rcmdr.
[Additional notes:] Recent version's Rcmdr supports the function to directly read *.xls files using RODBC library. Select [Data], [Import data], [from Excel, Access, or dBase data set], and enter an arbitrary dataset name into [Enter dataset name:] textbox, select an Excel book file to specify the sheet including the data.

2.2 Principle of data entry to avoid errors due to typos

The data entry should be duplicatedly done by more than two researchers. After completion of two files, those difference can be checked by some programs (for example, copying those two files' contents to separate worksheets of an Excel book, and entering the formula of

```
=If(Sheet1!A1=Sheet2!A1,"","X")
```

into corresponding cells of the third worksheet of the same book. If all cells of two files are same, the third sheet will show blank only. The cells in the third worksheet showing `X` must be checked with reference to raw data and fix them until achieving a looked-like blank third worksheet.

However, it is sometimes difficult to keep two researchers, double-entries by a researcher or comparing the printed-out data with screen are used instead.

2.3 How to treat missing values

It is necessary to pay attention to how to treat missing values. In general, the data to be statistically analyzed are sampled from the source population. The representativeness of the sample is necessary to draw a valid inference on the source population from the result of statistical analyses about the sample data. In both questionnaire (No Answer, Unknown, etc.) and experimental research (Below the detection limit, insufficient quantity of samples to measure, accidental loss of samples, etc.), how to treat missing values is critical point to add bias to sample representativeness.

For example, in a diet-related questionnaire, people who gave no answer to the question "Do you like sweets?" may like

sweets but didn't reply [Yes] because they had known too much sweets-intake being judged as harmful factor for health. If so, omitting them from the analysis cause to make bias: the sample may include less people who like sweets, compared with general population. The researcher must pay effort to reduce such missing values, and must pay attention to them in explaining the result of analyses.

The code for missing values is NA in R. It is blank in Excel, and blank field in the tab-delimited text file is read as NA in R.

How to treat the data including missing values has no golden rule. The most clear method is to exclude any case with missing values. If you do so, the easiest way to exclude the case with missing value is deleting such lines in Excel worksheet. However, if you have many cases with few missing values, you can leave missing values in the dataset, and exclude missing values in each analysis. Anyway, you should make effort to reduce missing values as possible as you can.

Considering more rigidly, although it is impossible to correct the bias caused by non-random missings as described above, we must consider two situations of random missings. For the cases of "MISSING COMPLETELY AT RANDOM (MCAR)", simply excluding the missing cases causes no bias but decreases statistical powers. However, for the cases of "MISSING AT RANDOM (MAR)" — there is no difference in the distributions of the variable between the observed cases and missing cases and there are significant differences in the distributions of the other variables —, simply exluding missing cases may cause bias. To avoid such biases, many methods of "multiple imputation" to compensate missing values were developed. In R, the two packages (`mitools`[*5] and `mice`[*6]) are famous. The maintainer of the latter package `mice` is Dr. Stef van Buuren, the specialist of multiplie imputation, who manages the web-site "Multiple Imputation Online"[*7]. Because it is complicated, I cannot give a detailed explanation here[*8].

2.4 Descriptive Statistics

The purposes to calculate descriptive statistis are, (1) glancing a feature of the data, and (2) checking the possibility of data entry error. Impossible maximum/minimum values or too large standard deviation suggest such data entry errors.

Descriptive statistics include the "central tendency" which shows the location of the data and the "variability" which shows the scale of the data.

The following 3 indices are popular central tendencies. Usually the mean is used, but the median is also used for the values with outliers or trimmed distribution.

mean Most frequently used location parameter of the distribution. The mean of the population μ (pronounced as "mu") is,

$$\mu = \frac{\left(\sum_{i=1}^{N} X_i\right)}{N}$$

X_i is each value in the distribution and N is total number of samples, where $\sum$ (pronounced as "sigma") means the sign of summation, *i.e.*,

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + X_3 + ... + X_N$$

The equation for sample mean is the same as the equation for population mean shown above. But the signs used in

[*5] `http://cran.r-project.org/web/packages/mitools/index.html`

[*6] `http://cran.r-project.org/web/packages/mice/index.html`

[*7] `http://www.multiple-imputation.com/`

[*8] For example, let the data frame with missing value `withmiss`. After loading the `mice` package by typing `library(mice)`, `imp <- mice(withmiss)` makes the object `imp`, which includes the original data with the coefficients to estimate missing values based on multiple imputation. The methods of multiple imputation can be selected from `"sample"`, `"pmm"`, `"logreg"`, `"norm"`, `"lda"`, `"mean"`, `"polr"`, and so on as the option `meth=` within the `mice()` function. To obtain the new data frame with compensated estimates for originally missing values, for example, type as `est <- complete(imp, 2)`. Here 2 means that the second coefficients set was used to estimate missing values among 5 (by default) coefficients sets. After that, multiple results based on multiple data frames with compensations must be compiled (integrated).

the equation are slightly different. The sample mean $\bar{X}$ (pronounced as "X bar") is defined as

$$\bar{X} = \frac{\left(\sum_{i=1}^{n} X_i\right)}{n}$$

where n is the sample size. Weighted mean is the summation of certain weights times values devided by the summation of weights. In equation, let the weights w_i,

$$\bar{X} = \frac{w_1\bar{X_1} + w_2\bar{X_2} + ... + w_n\bar{X_n}}{w_1 + w_2 + ... + w_n}$$

In Rgui console, `mean(X)` gives the mean of a numeric vector X.

median The median divides the whole data into larger half and smaller half. Calculation of median does not require equation but algorithm. From this nature, median hardly suffers from the effect of outliers. In Rgui console, `median(X)` gives the median of a numerical vector X.

mode The most frequently appearing value is the mode. In Rgui console, `table(X)[which.max(table(X))]` may give the mode (however, if there are some candidates with the same frequency, only the first one of them is given).

There are many other central tendencies like harmonic mean(= $1/(\sum_{i=1}^{n} \frac{1}{X_i})$), geometric mean (= $(\prod_{i=1}^{n} X_i)^{1/n}$). Both harmonic mean and geometric mean are less sensitive to outliers than mean, but these cannot be used data including zero.

The following 4 are the popular indices of variability.

Inter-Quartile Range; IQR The quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. Dividing ordered data into q essentially equal-sized data subsets is the motivation for q-quantiles; the quantiles are the data values marking the boundaries between consecutive subsets (See, Wikipedia). If $q = 4$, the quantile is called as quartile (If $q = 100$, the quantile is called as percentile). Quartiles is composed of 3 values: the first quartile [Q1], the second quartile [Q2], and the third quartile [Q3]. The Q2 is same as the median. The five values of 3 quartiles with minimum and maximum are called as five numbers, which is calculated by `fivenum()` function in R. The IQR is the interval between Q1 and Q3, which means central half of the distribution. In Rgui console, `IQR(X)` gives the IQR of a numeric vector `X`.

Semi Inter-Quartile Range; SIQR SIQR is IQR/2. If the data obeys normal distribution, central half of the data is included from the median minus SIQR to the median plus SIQR. SIQR hardly suffers from outliers.

variance Deviation is the difference between each value and the mean. To equally treat minus deviation and plus deviation, we can think the mean of squared deviation, that is the variance. The variance V is defined as $V = \frac{\sum(X-\mu)^2}{N}$, actually $V = \frac{\sum X^2}{N} - \mu^2$. As sample variance, instead of dividing squared deviance by the sample size n, dividing by $n-1$, that is called as the unbiased variance. The unbiased variance is a better estimate of the population variance than original variance. The unbiased variance V_{ub} is defined as $V_{ub} = \frac{\sum(X-\bar{X})^2}{n-1}$. In Rgui console, `var(X)` gives the unbiased variance of a numeric vector `X`.

standard deviation; sd The standard deviation is square-root of variance, **to match the dimention with the mean**. The unbiased standard deviation is square-root of unbiased variance[*9]. If the distribution of the data is normal distribution, the interval between the mean minus 2 sd and the mean plus 2 sd[*10] includes approximately 95% of the data. In Rgui console, `sd(X)` gives the unbiased standard deviation of a numeric vector `X`.

In Rcmdr, select [Statistics], [Summaries], [Numerical Summaries], then you can get mean, standard deviation, minimum, first quartile, median, third quartile, maximum, and the number of samples. Please try this using the previously entered data set about 10 subjects' height and weight.

[*9] To note, the unbiased variance is the unbiased estimate of the population variance, but the unbiased standard deviation is not the unbiased estimate of the population standard deviation. Here "unbiased" just means the source of calculation being the unbiased variance.

[*10] Usually it is written as mean $\pm 2sd$. The value 2 is approximation of 97.5 percent point of the standard normal distribution, 1.95996398454...: You can see it by typing `options(digits=20)` and `qnorm(0.975)` in Rgui console.

2.5 Drawing Figures

To capture the whole nature of the data, I recommend to draw graphs. Human ability of visual perception is superior to computer, at least about pattern recognition. Drawing graphs is also effective to find data entry errors.

How to draw suitable graphs depends on the scale of variables. For discrete variables, popular graphs are: Frequency bar plot, stacked bar plot, horizontal bar plot, pie chart, and so on.

Let's try the example. Drawing graphs for discrete variables by Rcmdr, the variables should have "factor" attributes. So we should change the active dataset to "survey" included in "MASS" package. At first, select [Tools], [Load package(s)...], then select "MASS" and click [OK]. Next, select [Data], [Data in packages], [Read data set from an attached package], then double-click [MASS] in the left box, double-click [survey] in the right box, and click [OK], sequentially.

The data frame "survey" in MASS library contains the responses of 237 students at the University of Adelaide to a number of questions (Venables and Ripley, 1999). The variables are:

Sex The sex of the student. (Factor with 2 levels "Male" and "Female".)

Wr.Hnd The span (distance from tip of thumb to tip of little finger of spread hand) of writing hand, in centimetres.

NW.Hnd The span of non-writing hand.

W.Hnd Writing hand of student. (Factor, with 2 levels "Left" and "Right".)

Fold The answer to "Fold your arms! Which is on top?" (Factor, with 3 levels "R on L", "L on R", "Neither".)

Pulse Pulse rate of student (beats per minute).

Clap The answer to "Clap your hands! Which hand is on top?" (Factor, with 3 levels "Right", "Left", "Neither".)

Exer How often the student exercises. (Factor, with 3 levels "Freq" (frequently), "Some", "None".)

Smoke How much the student smokes. (Factor, with 4 levels "Heavy", "Regul" (regularly), "Occas" (occasionally), "Never".)

Height Height of the student in centimetres.

M.I Whether the student expressed height in imperial (feet/inches) or metric (centimetres/metres) units. (Factor, with 2 levels "Metric", "Imperial".)

Age Age of the student in years.

Using this data set, let's draw several graphs.

Frequency bar plot To draw the frequency of each category as vertical bars, by categories. For example, drawing the frequency bar plot for the variable `Smoke` in `survey`, in Rgui console, `barplot(table(survey$Smoke))`.

In Rcmdr, select [Graphs], [Bar Graph], then select [Smoke] and click [OK]. Alignment of bars usually obeys the alphabetical order of the category names (It can be changed using [Data], [Manage variables in active data set], [Reorder factor levels...]).

Stacked bar plot The graph with stacked bars. It can be drawn by typing in Rgui console as follows.

```
barplot(as.matrix(table(survey$Smoke)))
```

It is not supported in Rcmdr.

Horizontal bar plot Horizontally stacked bars with percentages. It can be drawn by typing in Rgui console as follows.

```
barplot(as.matrix(table(survey$Smoke)/NROW(survey$Smoke)*100),horiz=TRUE)
```

It is also not supported in Rcmdr.

Pie chart Sectored circle due to proportions of categories. In Rgui console, `pie(table(survey$Smoke))`. This graph is well known but not recommended by R Development Core Team, because the human eyes are good at judging linear measures and bad at judging relative areas. They recommend a bar plot or dot chart for displaying this type of data.

In Rcmdr, select [Graphs], [Pie chart...], then select [Smoke] and click [OK].

For continuous variables, the polular graphs are the followings.

Histogram To see the distribution of a single numeric variable, plot the counts in the porperly spaced cells defined by "breaks". By default, breaks are calculated using "Sturges" algorithm, but it can be given explicitly. And by default, the cells are intervals of the form "(a, b]". If you need "[a, b)", `right=FALSE` option must be specified (this option is not supported in Rcmdr). To draw a histogram of `Age` in `survey` data set (the range of `Age` is from 16.75 to 73), type `hist(survey$Age)`. If you want to define the cells as [10,20), [20, 30), ..., [70, 80), type `hist(survey$Age, breaks=1:8*10, right=FALSE)`.

In Rcmdr, select [Graphs], [Histogram...], then select [Age] and click [OK]. Some options can be selected.

Normal QQ plot To see whether the distribution of a single numeric variable is normal distribution or not, the data points are plotted to corresponding quantiles of a normal distribution: If the data obey a normal distribution, the graph looks on a straight line. To draw this for `Pulse` in `survey` data set in Rgui console, simply type `qqnorm(survey$Pulse)`.

In Rcmdr, select [Graphs], [Quantile comparison plot...], then select [Pulse] and click [OK]. Some options can be selected.

Stem and leaf plot Stacking rough value as stem and aligning the lowest digits as leafs. The whole shape is similar to histogram, but plotting numbers instead of rectangles. In Rgui console, type `stem(survey$Pulse)`.

In Rcmdr, select [Graphs], [Stem-and-leaf display...], then select [Pulse] and click [OK]. The stem and leaf plot is drawn in the output window, instead of graph window.

Box and whisker plot Draw a box with top line being Q3 and bottom line being Q1, where the center line is median. Adding 1.5 times IQR "whisker" on top and bottom, but if the whiskers go over minimum or maximum, it must be cut there. If there are outliers beyond the whiskers, those will be plotted as small circles. Stratified box-and-whisker plot is useful to compare the distributions among strata. For example, to draw the box-and-whisker plot of `Pulse` stratified by `Smoke` in Rgui console, type `boxplot(survey$Pulse ~ survey$Smoke)`.

In Rcmdr, select [Graphs], [Boxplot...], then select [Pulse] and click [Plot by groups...] and select [Smoke] and click [OK], and again click [OK]. As similar graph, plotting means with error bars is also possible. Select [Graphs], [Plot of means...], then select [Smoke] as Factors (left box) and [Pulse] as Response Variable (right box). After checking the kind of error bars (standard error, standard deviation, and confidence intervals are possible), click [OK].

Radar chart More than 3 variables are radially aligned and connecting plots as polygons. It is also known as spider chart. One radar chart will be made for each subject, so that the comparison of several radar charts will become possible by drawing multiple charts in one figure. To draw this, additional package (`plotrix` or `fmsb`) is needed. From CRAN mirror sites, you can download and install them by typing `install.packages("plotrix")` and `install.packages("fmsb")`. Once doing so, type `library(fmsb)` and `example(radarchart)` in Rgui console, you will know how to use it. **This graph is not supported by Rcmdr.**

Scatter plot To show relationships between 2 continuous variables, plotting the points with one variable as x axis and the other as y axis. For example, you can see the relationships between height `Height` as y-axis and age `Age` as x-axis in `survey` data set, by typing `plot(Height ~ Age, data=survey)` in Rgui console. If you plot the points of Males and Females in different color/mark, you can use `pch=as.integer(Sex)` and `col=c("Pink","Blue")[as.integer(Sex)]` options.

In Rcmdr, select [Graphs], [Scatterplot...], then select [Age] as x-variable (left box) and [Height] as y-variable (right box), and click [OK]. Several options can be specified, including stratified plotting. If you would like to identify each data points by clicking graph, before clicking [OK], check the box of [Identify points].

3 Statistical tests to compare 2 groups

Medical research has traditionally preferred "hypothesis testing". But, the hypothesis testing is to limit the information originally included in data into the simple binary information whether the hypothesis can be rejected or not. It's too simplifying, but traditionally used in many publications. Nonetheless, some modern epidemiologists like Kenneth J. Rothman or Sandra Greenland recommend to use the estimation of confidence intervals or drawing p-value plot[*11], instead of hypothesis testing.

As a typical example, let's see the testing of null-hypothesis that the means of independent 2 groups are not different. Usually the researcher must determine the significance level of the test in advance. The significance level of a test is such that the probability of mistakenly rejecting the null hypothesis is no more than the stated probability. There are two ways of thinking. In a Fisherian manner, the **p-value** (significant probability) is the probability conditional on the null hypothesis of the observed data or more extreme data. If the obtained p-value is small then it can be said either the null hypothesis is false or an unusual event has occurred. In a Neyman-Pearson's manner, both a null and an alternative hypothesis must be defined and the researcher investigates the repeat sampling properties of the procedure, *i.e.* the probability that a decision to reject the null hypothesis will be made when it is in fact true and should not have been rejected (this is called a "false positive" or Type I error) and the probability that a decision will be made to accept the null hypothesis when it is in fact false (Type II error). These two way of thinking should be distinguished.

Usually, the significance level should be set as 0.01 or 0.05 before the hypothesis testing and if the obtained p-value is less than the significance level, the null-hypothesis is rejected to judge that there is a statistical significance.

A summary of statistical hypothesis testing between independent 2 groups is:

1. Continuous variable:
   (a) Obeying normal distribution[*12]: Welch's *t*-test (in Rgui console, `t.test(x,y)`)[*13]
   (b) Otherwise: Wilcoxon's rank sum test (in Rgui console, `wilcox.test(x,y)`)
2. Categorical variable: chi-square test for proportions (in Rgui console, `prop.test()`)

[*11] This function is implemented as `pvalueplot()` function in the `fmsb` package previously mentioned.

[*12] It can be tested by Shapiro-Wilk test (using `shapiro.test()` in Rgui console, but it is not recommended to simply apply the result to determine whether the non-parametric test might be used or not.

[*13] Some researchers recommend to do *F*-test for the null-hypothesis of equal variances in variance, and if the variances are different, those two samples might come from different populations. Traditionally, the two-stage testing was recommended by some statisticians, that is, doing normal *t*-test if the result of *F*-test is not significant and Welch's *t*-test otherwise. But recently according to the simulation studies by Dr. Shigenobu Aoki (Gunma Univ.), it was proved that Welch's *t*-test can achieve the most unbiased result compared with such two-stage testing. So Welch's *t*-test is always recommended.

3.1 F-test for the testing equal variances

Assume two continuous variables `X` and `Y`. Calculate the unbiased variances of these two variables, `SX<-var(X)` and `SY<-var(Y)`. If `SX>SY`, calculate the ratio of the larger to smaller, as `F0<-SX/SY`. The `F0` obeys the F-distribution of the first degree of freedom (d.f.) `DFX<-length(X)-1` and the second d.f. `DFY<-length(Y)-1`. Therefore, the p-value is `1-pf(F0,DFX,DFY)`. Simply, `var.test(X, Y)` can do so. If the data frame includes one quantitative variable `X` and one group variable `C`, it can be done by `var.test(X~C)`. For example, to test the null-hypothesis that the variances of heights (`Height`) are not different by sex (`Sex`) in the `survey` data set, type `var.test(Height ~ Sex, data=survey)`.

In Rcmdr, select [Statistics], [Variances], [Two variances F-test], then select [Sex] as [Groups] (left box) and [Height] as [Response Variable] (right box), and click [OK]. To be shown as the candidates of [Groups], the variable must be a factor. If the variable to be used as [Groups] is numeric, it can be changed using [Data], [Manage variables in active data set], [Convert numeric variables to factors].

3.2 Welch's t-test for the testing equal means

Calculate $t_0 = |E(X) - E(Y)|/\sqrt{S_X/n_X + S_Y/n_Y}$. The t_0 obeys t-distribution of the degree of freedom ϕ, where

$$\phi = \frac{(S_X/n_X + S_Y/n_Y)^2}{\{(S_X/n_X)^2/(n_X-1) + (S_Y/n_Y)^2/(n_Y-1)\}}$$

In Rgui console, simply type `t.test(X, Y)` or `t.test(X~C)`. For example, to test the null hypothesis that the mean heights (`Height`) are not different by sex (`Sex`) in the `survey` data set, type `t.test(Height ~ Sex, data=survey)`.

In Rcmdr, select [Statistics], [Means], [Independent samples t-test], then select `Sex` as [Groups] and `Height` as [Response Variable] and click [OK]. The result will appear in the Output Window.

When you have means and unbiased standard deviations for 2 groups, popular expression of the graph is barplot with error bars[*14], but if you have raw data, the stripchart will be usually drawn.

An example. If there are the 2 numerical variables `V <- rnorm(100,10,2)` and `W <- rnorm(60,12,3)`, those can be converted as follows.

```
X <- c(V,W)
C <- as.factor(c(rep("V",length(V)),rep("W",length(W))))
x <- data.frame(X,C)
```

or

```
x <- stack(list(V=V,W=W))
names(x) <- c("X","C")
```

Then we can make stripcharts with error bars as follows.

```
stripchart(X~C, data=x, method="jitter", vert=TRUE)
Mx <- tapply(x$X,x$C,mean)
Sx <- tapply(x$X,x$C,sd)
Ix <- c(1.1,2.1)
points(Ix, Mx, pch=18, cex=2)
arrows(Ix, Mx-Sx, Ix, Mx+Sx, angle=90, code=3)
```

[*14] `barplot()` and `arrows()` will be used to draw this kind of graph.

3.3 Paired *t*-test

If the comparison is done between 2 paired values for each subject (for example, comparison between the before and after the treatment), paired *t*-test is more effective than independent two sample *t*-test, because it considers individual differences.

The paired *t*-test is exactly same as (1) to calculate each difference and (2) test the null-hypothesis that mean difference is not different from zero. For example, to test the two hand sizes by the paired *t*-test in `survey`, type `t.test(survey$NW.Hnd, survey$Wr.Hnd, paired=TRUE)` or `t.test(survey$NW.Hnd-survey$Wr.Hnd,mu=0)`. To draw appropriate graph, type as below.

```
Diff.Hnd <- survey$Wr.Hnd - survey$NW.Hnd
C.Hnd <- ifelse(abs(Diff.Hnd)<1,1,ifelse(Diff.Hnd>0,2,3))
matplot(rbind(survey$Wr.Hnd, survey$NW.Hnd), type="l", lty=1, col=C.Hnd, xaxt="n")
axis(1,1:2,c("Wr.Hnd","NW.Hnd"))
```

In Rcmdr, select [Statistics], [Means], [Paired t-test], then select [NW.Hnd] as [First variable] and [Wr.Hnd] as [Second variable] and click [OK].

Exercise

In Rcmdr, select [Data], [Data in packages], [Read data set from an attached package...], then double-click `datasets` from the left panel, then double-click `infert` from the right panel, then click [OK]. You may successfully load the "infert" data (cited from Trichopoulos *et al.* (1976) Induced abortion ans secondary infertility. *Br J Obst Gynaec*, 83: 645-650).

The data include several variables from the OB/GYN patients with secondary infertility, of whom original candidates were 100, but two controls with matched age, parity and education were found for only 83 patients, so that the number of samples were 248 (because the 74th patients had only one matched control, another control was excluded because who had each two times of spontaneous and induced abortions, respectively).

Included variables are:

education Factor variable to show education period, with 3 levels: 0 = "0-5 years", 1 = "6-11 years", 2 = "12+ years"

age Numeric variable for age in years of case

parity Numeric variable for the number of ever-borne children

induced Numeric variable for the number of prior induced abortions with 2 values: 1 = 1, 2 = 2 or more.

case Numeric variable to show the status of case or control: 1 means case, 0 means control.

spontaneous Numeric variable to show the number of prior spontaneous abortions: 0 = 0, 1 = 1, and 2 = 2 or more.

stratum Integer variable to show the matched set number: 1-83.

pooled.stratum Numeric variable to show the pooled stratum number: 1-63.

(1) Test the null-hypothesis that the mean numbers of prior spontaneous abortions are not different betweem case and control. (2) Test the null-hypothesis that the mean number of induced abortion is not different from the mean number of spontaneous abortion for each female. In both test, let the significance level 0.05.

This data set is much more suitable for the fitting of the logistic regression model, but here we try to check the differences of means. In this data set, "2 or more" is coded as 2, so that the exact means cannot be calculated, but here we ignore this incorrectness.

By Rgui console, it's very simple. To executing required *t*-test, type the following lines.

(1) `t.test(spontaneous ~ case, data=infert)`[*15]

(2) `t.test(infert$induced, infert$spontaneous, paired=TRUE)`

[*15] And `var.test(spontaneous ~ case, data=infert)` if you would like to test the null hypothesis that the variances of cases and controls are equal.

In Rcmdr, group variable must be set as Factor. Therefore, select [Data], [Manage variables in active data set], [Convert numeric variables to factors], then select `case` and type `group` in the box named as "New variable name or prefix for multiple variables:" and click [OK]. In the window to specify the level names for groups, type `control` in the box of `0` and type `infertile` in the box of `1` and click [OK].
After that, select [Graphs], [Boxplot], then select `spontaneous` and click [Plot by groups...] and select `group`, then click [OK] and click [OK]. By doing so, you can graphically see the box and whisker plots for controls and infertile groups separately. Otherwise, select [Graphs], [Plot of means...], then select `group` as [Factors] in the left box and `spontaneous` as [Response Variable] in the right box, and check the box beside [Standard deviations], and click [OK]. You will see the plot of means (connected by solid line) with error bars of unbiased standard deviations.
Then, to answer (1), select [Statistics], [Means], [Independent samples t-test], then select `group` as [Groups] and `spontaneous` as [Response Variable] and checking the radio button beside [No] of [Assume equal variances?], then click [OK]. The result will appear in the Output Window (If you would also like to test the equal variance hypothesis, select [Statistics], [Variances], [Two-variances F-test...], then select `group` as [Groups] and `spontaneous` as [Response Variable], and click [OK]).
To answer (2), select [Statistics], [Means], [Paired t-test...], then select `spontaneous` as [First variable] and `induced` as [Second variable], then click [OK].

3.4 Wilcoxon's rank sum test

Wilcoxon's rank sum test is an typical nonparametric test to compare the location parameter of 2 independent groups, which corresponds to *t*-test to compare the means of two independent groups. It is mathematically equivalent test with Mann-Whitney's *U*-test.

The principle of the Wilcoxon's rank sum test is to compare ranks instead of quantitative values. The procedure can be summarized as follows.

1. Let a variable X contain the values of $\{x_1, x_2, ..., x_m\}$ and another variable Y contain the values of $\{y_1, y_2, ..., y_n\}$.
2. First of all, mix the all values of X and Y and rank them in ascending order[*16]. For example, $x_8[1], y_2[2], y_{17}[3], ..., x_4[N]$, where $N = m + n$.
3. Next, calculating the sum of ranks for each variable. However, the overall sum of ranks is clearly $(N + 1)N/2$, so that calculate the sum of ranks for X. The sum of ranks for Y can be calculated as those difference.
4. Let the rank of x_i $(i = 1, 2, ..., m)$ included in X as R_i, the sum of rank for X is

$$R_X = \sum_{i=1}^{m} R_i$$

   Here, too large or too small R_X suggests that the null hypothesis "$H0$: The location of distribution X and that of Y are not different." is improbable[*17].
5. Under the null-hypothesis, X is randomly extracted m samples from N samples, and the rest consititutes Y. About the rank, extract m numbers from the rank of $1, 2, 3, ..., N$. Ignoring ties, the number of possible combinations are ${}_NC_m$[*18].
6. If $X > Y$, let the number of cases that the sum of ranks is equal to or larger than R_X k, among ${}_NC_m$.
7. If $k/{}_NC_m < \alpha$ (α is the significance level), reject $H0$. That's the way of exact calculation.

[*16] If the values contain ties, special treatment is needed.
[*17] If we write the sum of rank for X as `RS`, `2*(1-pwilcox(RS,m,n))` gives the exact p-value of two-sided Wilcoxon's rank sum test of the null-hypothesis in Rgui console.
[*18] `choose(N,m)` in Rgui console.

8. For large samples, normal approximation is applicable. Under the null-hypothesis, the expected value of rank sum $E(R)$ is, because each rank value may be at equal probability from 1 to N,

$$E(R) = \sum_{i=1}^{m} E(R_i) = m(1 + 2 + ... + N)/N = m(N + 1)/2$$

The variance of rank sum is, because $var(R) = E(R^2) - (E(R)^2)$,

$$E(R^2) = E((\sum_{i=1}^{m} R_i)^2) = \sum_{i=1}^{m} E(R_i^2) + 2\sum_{i<j} E(R_i R_j)$$

$$E(R_i^2) = (1^2 + 2^2 + ... + N^2)/N = (N + 1)(2N + 1)/6$$

$$E(R_i R_j) = \frac{1}{N(N-1)}\{(\sum_{k=1}^{N} k)^2 - \sum_{k=1}^{N} k\}$$

$$= \frac{1}{N(N-1)}(\frac{N^2(N+1)^2}{4} - \frac{N(N+1)(2N+1)}{6})$$

$$= \frac{(N+1)(3N+2)}{12}$$

Then finally we obtain:

$$var(R_X) = m(N + 1)(N - m)/12 = mn(N + 1)/12$$

9. From expected value and variance, standardize[*19] the rank sum with continuity correction[*20], then caluculate,

$$z_0 = \{|R_X - E(R_X)| - 1/2\}/\sqrt{var(R_X)}$$

For large m and n, z_0 obeys the standard normal distribution, so that we can judge the result showing statistically significant difference at 5% significance level if $z_0 > 1.96$. In Rgui console, let z_0 as `z0`, simply typing `2*(1-pnorm(z0,0,1))` gives the p-value.

10. However, special treatment is necessary for ties. For example, the variable X is $\{2, 6, 3, 5\}$ and Y is $\{4, 7, 3, 1\}$, the value 3 is included in both X and Y. In such case, giving mean rank (as shown in the table below) to both leads to solution.

Variable	Y	X	X	Y	Y	X	X	Y
Value	1	2	3	3	4	5	6	7
Rank	1	2	3.5	3.5	5	6	7	8

11. Nonetheless, the variance of rank sum, used in the normal approximation, will change. Under the null-hypothesis,

$$E(R_X) = m(N + 1)/2$$

is unchanged but

$$var(R_X) = mn(N + 1)/12 - mn/\{12N(N - 1)\} \cdot \sum_{t=1}^{T}(d_t^3 - d_t)$$

where T is the number of sets of ties and d_t is the number of tth tied data. In the above example, $T = 1$ and $d_1 = 2$.

To practice, activate `survey` data set again. Let's test the null-hypothesis that the locations of the distribution of `Height` are not different by `Sex`.

In Rgui console, it can be done by `wilcox.test(Height ~ Sex, data=survey)`.

[*19] Subtracting mean from each value and dividing them by square root of the variance.

[*20] To improve approximation by the normal distribution, subtracting or adding 1/2 about each number.

In Rcmdr, after selecting [survey] as active [Data set], select [Statistics], [Nonparametric tests], [Two-sample Wilcoxon test...], then select `Sex` in the [Groups] box and `Height` in the [Response Variable] box, and click [OK]. Here you can explicitly specify the type of test as exact test, normal distribution, or normal approximation with continuity correction.

3.5 Wilcoxon's signed rank test

Wilcoxon's signed rank test is the nonparametric version of paired *t*-test. Here I will not give any explanation, but it is explained in many statistics textbooks.

To practice, let's test the null-hypothesis that the locations of the distribution of `Wr.Hnd` and `NW.Hnd` of `survey` data set are not different. Here we also set the significance level as 5%.

In Rgui console, simply typing `wilcox.test(survey$Wr.Hnd, survey$NW.Hnd, paired=TRUE)` gives the result. If the `p-value` shown in the Output Window is less than 0.05, we can judge to reject the null-hypothesis.

In Rcmdr, select [Statistics], [Nonparametric tests], [Paired-samples Wilcoxon test...], then select `Wr.Hnd` as the [First variable] at the left box and `NW.Hnd` as the [Second variable] at the right box, and click [OK]. You can specify the type of test as same as the case of Wilcoxon's rank sum test, though usually leaving that radio button [Default] is enough.

3.6 Testing the equality of proportions in two independent groups

Let's consider the n_1 patients and n_2 controls. The numbers of individuals with a specific feature were r_1 in the patients and r_2 in the controls, then the sample proportions with the feature are $\hat{p_1} = r_1/n_1$ in the patients and $\hat{p_2} = r_2/n_2$ in the controls.

Here we will test the null-hypothesis that those proportions in the patients' source population and in the controls' source population (p_1 and p_2, respectively) are not different.

The point estimate of p_1 is $\hat{p_1}$ and the point estimate of p_2 is $hat p_2$. The expected difference between p_1 and p_2 can be estimated as $\hat{p_1} - \hat{p_2}$ and the variance of the difference can be calculated as $V(\hat{p_1} - \hat{p_2}) = p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2$. Under the null-hypothesis, we can assume $p_1 = p_2 = p$, then $V(\hat{p_1} - \hat{p_2}) = p(1 - p)(1/n_1 + 1/n_2)$. Replacing p by $\hat{p} = (r_1 + r_2)/(n_1 + n_2)$, and denoting $\hat{q} = 1 - \hat{p}$, when $n_1 p_1 > 5$ and $n_2 p_2 > 5$, we can standardize $\hat{p_1} - \hat{p_2}$ and apply the normal appoximation,

$$Z = \frac{\hat{p_1} - \hat{p_2} - E(\hat{p_1} - \hat{p_2})}{\sqrt{V(\hat{p_1} - \hat{p_2})}} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}} \sim N(0, 1)$$

For the continuity correction,

$$Z = \frac{|\hat{p_1} - \hat{p_2}| - (1/n_1 + 1/n_2)/2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

and we can reject the null-hypothesis when Z is greater than 1.96.

For example, let's test the null-hypothesis that the smoker's proportions are not differenct between both 100 patients and controls where the numbers of smokers were 40 and 20, respectively.

```
p <- (40+20)/(100+100)
q <- 1-p
Z <- (abs(40/100-20/100)-(1/100+1/100)/2)/sqrt(p*q*(1/100+1/100))
2*(1-pnorm(Z))
```

Typing above into Rgui console, `[1] 0.003370431` will be obtained. Because 0.0033... is less than 0.05, we can judge to reject the null-hypothesis at 5% significance level.

The 95% confidence intervals of the expected difference of proportions can be estimated by typing the following.

```
dif <- 40/100-20/100
vardif <- 40/100*(1-40/100)/100+20/100*(1-20/100)/100
difL <- dif - qnorm(0.975)*sqrt(vardif)
difU <- dif + qnorm(0.975)*sqrt(vardif)
cat("Expected difference of proportions=",dif,"  95% conf.int.= [",difL,",",difU,"]\n")
```

The result will be [0.076,0.324]. For the continuity correction, subtracting $(1/n_1 + 1/n_2)/2 = (1/100 + 1/100)/2 = 0.01$ from the lower limit and adding the same value to the upper limit, the 95% confidence interval will be [0.066,0.334].

In Rgui console, those calculation can be done by simply typing as follows.

```
smoker <- c(40,20)
pop <- c(100,100)
prop.test(smoker,pop)
```

To practice, let's consider **the null-hypothesis that the proportions of lefty writer are not different between males and females** in `survey` data set.

In Rgui console, typing `prop.test(table(survey$Sex,survey$W.Hnd))` gives result of 2-sample test for equaility of proportions with continuity correction. The `prop 1` is the proportion of lefty writers in females and the `prop 2` is that in males. To judge the statistical significance, see the `p-value`. **Actually the test of proportions by normal approximation is mathematically equivalent to the test of chi-square test of contingency table, so that typing** `chisq.test(table(survey$Sex,survey$W.Hnd))` **gives the same** `p-value`**.**

In Rcmdr, select [Statistics], [Proportions], [Two-sample proportions test...], then select `Sex` as [Groups] and `W.Hnd` as [Response Variable], and check the radio button beside [Normal approximation with continuity correction] as the [Type of Test], and click [OK].

To obtain `p-value`, select [Statistics], [Contingency tables], [Two-way table], then select `Sex` as [Row variable] and `W.Hnd` as [Column variable], and click [OK]. The resulted `p-value` is calculated without continuity correction. In the current version of Rcmdr, there is no option to specify using continuity correction in this testing.

4 Testing the difference of locations among 3 or more groups

To compare the means among 3 or more groups, you must not simply repeat *t*-tests for all possible pairs. In the statistical hypothesis testing, setting significance level as 5% in each comparison provides much more type I error for overall comparisons.

To solve this problem, there are two different approaches. (1) Evaluate the effect of a group variable on a quantitative variable. (2) Adjust the type I errors for multiple comparisons. Traditionally these approaches are conducted as two-steps: Only when the effect of the group is statistically significant, the pairwise comparisons with adjustment for multiple comparisons will be done. In that meaning, the latter step is called as *post hoc* analysis of the former step. However, it is recommended now that you should select more appropriate approach according to the purpose of the analysis.

4.1 One-way Analysis of Variance (ANOVA)

The typical analysis of the former approach to test the means among 3 or more groupps is one-way analysis of variance (ANOVA). The one-way ANOVA decomposes the variance of the data into the variance by the group and the variance of

random errors. If the ratio of these variance is significantly different from 1, we can judge the effect of the group variable is statistically significant.

For example, the heights of 37 adult males in 3 villages in the South Pacific were as follows. You can read the data from the web site[*21].

Unique ID (PID)	Village (VG)	Height (cm) (HEIGHT)	Weight (kg) (WEIGHT)
10101	X	161.5	49.2
10201	X	167.0	72.8
⋮			
30301	Z	166.0	58.0
⋮			
70312	Y	155.5	53.6

In Rgui console, to conduct one-way ANOVA of the null-hypothesis that VG has no significant effect on HEIGHT, type as follows.

```
sp <- read.delim("http://phi.med.gunma-u.ac.jp/grad/sample2.dat")
summary(aov(HEIGHT ~ VG, data=sp))
```

Then you get the following result, so-called "ANOVA summary table".

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
VG           2  422.72  211.36  5.7777 0.006918 **
Residuals   34 1243.80   36.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here the number of * at right is the codes for significance shown in the bottom line, but the Pr(>F), the significant probability, is more important. The column of Sum Sq means the sum of squared difference of each value with the mean. The value of VG's Sum Sq is 422.72, which is the sum of number of individuals times the squared difference between the mean in each village and the overall mean. This value is called as inter-group (inter-class) variation. The Residuals's Sum Sq, 1243.80, is the sum of squared difference between each individual's height and the mean height of the village where the individual belongs to, which is called as within-group (within-class) variation or error variation. The Mean Sq is mean squared difference, which is the value of Sum Sq divided by Df, where Df is degree of freedom. The Mean Sq is the variance, so that VG's Mean Sq, 211.36 is the inter-group variance and Residuals's Mean Sq, 36.58 is the error variance. Then, the F value is the ratio of variances, which is the value of inter-group variance divided by error variance. The ratio of variances obeys F-distribution, The significant probability Pr(>F) is obtained from F-distribution. Here the Pr(>F) is 0.006918, so that the effect of VG can be judged as statistically significant at 5% significance level. We can reject the null-hypothesis, and the heights of them significantly differ by village.

In Rcmdr, to read sample2.dat into sp data set, select [Data], [Import data], [from text file, clipboard, or URL...], then type sp **in the box of "Enter name for data set:", check the ratio button beside "Internet URL", and check the radio button beside [Tabs] of "Field Separator" and click [OK], then type** http://phi.med.gunma-u.ac.jp/grad/sample2.dat **as Internet URL and click [OK]. Next, select [Statistics], [Means], [One-way ANOVA...], then select** VG **as [Groups] and** HEIGHT **as [Response Variable], and click [OK].**

Traditionally speaking, one-way ANOVA should be done only after testing the null-hypothesis that the vari-

[*21] http://phi.med.gunma-u.ac.jp/grad/sample2.dat

ances of groups are not different (by Bartlett's test or other test). In Rgui console, you can do so by typing `bartlett.test(HEIGHT ~ VG, data=sp)`. The obtained p-value, 0.5785 means that we cannot reject null-hypothesis. After confirming it, we can safely apply one-way ANOVA.

In Rcmdr, select [Statistics], [Variances], [Bartlett's test...], then select `VG` as [Groups] and `HEIGHT` as [Response Variable] and click [OK].

However, such kind of two-steps analysis may cause multiple comparisons problem, so that Welch's extended one-way ANOVA without testing equal variance is the most appropriate.

In Rgui console, type `oneway.test(HEIGHT ~ VG, data=sp)`, then you can get the p-value by the Welch's extended one-way ANOVA. The p-value, 0.004002 is less than 0.05, so that we can judge the effect of village on the height is statistically significant at 5% significance level. Unfortunately, Rcmdr does not support `oneway.test()` now.

4.2 Kruskal-Wallis test and Fligner-Killeen test

The Kruskal-Wallis test is the nonparametric test to compare the locations of distributions among 3 or more groups, so that it seems a nonparametric alternative of ANOVA.

In Rgui console, type `kruskal.test(HEIGHT ~ VG, data=sp)`.

In Rcmdr, select [Statistics], [Nonparametric tests], [Kruskal-Wallis test...], then select `VG` as [Groups] and `HEIGHT` as [Response Variable], and click [OK].

The Fligner-Killeen test is the test the null hypothesis that the variances in each of the groups are the same in nonparametric manner. It seems a nonparametric alternative of Bartlett's test.

In Rgui console, type `fligner.test(HEIGHT ~ VG, data=sp)`. It is not supported in Rcmdr now.

4.3 Pairwise comparisons with adjustment of multiple comparisons

There are many methods to adjust type I errors such as Bonferroni's method, Holm's method, Tukey Honest Significant Differences (Tukey's HSD), and so on. Except Tukey's HSD, many adjustment methods are applicable in `pairwise.t.test()`, `pairwise.wilcox.test()`, or `pairwise.prop.test()`, where the `pairwise.prop.test()` is to test the proportions of 3 or more groups.

To compare means, pairwise *t*-test with Holm's adjustment or Tukey's HSD can be used. To compare medians, pairwise Wilcoxon's rank sum test with Holm's adjustment is appropriate.

For example, compare the means of height between the all pairs among 3 villages in `sp` data set.

In Rgui console, type either of the following lines:

```
pairwise.t.test(sp$HEIGHT, sp$VG, p.adjust.method="holm")
TukeyHSD(aov(HEIGHT ~ VG, data=sp))
```

In Rcmdr, only Tukey's HSD is applicable. Select [Statistics], [Means], [One-way ANOVA...], then select `VG` as [Groups] and `HEIGHT` as [Response Variable] and check the box beside "Pairwise comparisons of means", then click [OK]. After the result of ANOVA, the result of Tukey's HSD will appear in the Output Window and the graph of simultaneous confidence intervals for each group.

4.4 Dunnett's multiple comparisons

Dunnett's method to adjust multiple comparison of means is used in the comparison of multiple treatment groups with a common control group. For example, randomly assign the 15 hypertension patients to 3 groups (each 5), which are treated by placebo, by usual drug, and by new drug. After 1 month treatment, the decrease of systolic blood pressures (mmHg) were [5, 8, 3, 10, 15] for the placebo group, [20, 12, 30, 16, 24] for the usual drug group, and [31, 25, 17, 40, 23] for the new drug group.

Here we can apply the Dunnett's multiple comarisons to compare the mean reductions of systolic blood pressure of placebo group with both usual drug group and new drug group. In Rgui console, type below. Unfortunately, it is not supported in Rcmdr.

```
bpdown <- data.frame(
 medicine=factor(c(rep(1,5),rep(2,5),rep(3,5)),
 labels=c("placebo","usualdrug","newdrug")),
 sbpchange=c(5, 8, 3, 10, 15, 20, 12, 30, 16, 24, 31, 25, 17, 40, 23))
summary(res1 <- aov(sbpchange ~ medicine, data=bpdown))
library(multcomp)
res2 <- glht(res1, linfct = mcp(medicine = "Dunnett"))
confint(res2, level=0.95)
summary(res2)
```

5 Testing the differences of proportions among 3 or more groups

The R function of `prop.test()` is applicable for the comparison among 3 or more groups, where the null-hypothesis is that there is no difference of proportions among all groups. If you can reject the null-hypothesis, usually do pairwise comparison with adjustment of multiple comparisons.

Let's consider the `survey` data set in `MASS` package again. To compare the lefty proportions among the 3 groups of clapping hands with pairwise comparisons, type into Rgui console as follows.

```
prop.test(table(survey$Clap, survey$W.Hnd))
pairwise.prop.test(table(survey$Clap, survey$W.Hnd), p.adjust.method="holm")
```

In Rcmdr, we cannot simply compare proportions among 3 or more groups. However, as already shown, the chi-square test of contingency table is mathematically equivalent with this, so that it can be done. After activating `survey` (by clicking the box beside [Data set:] and selecting `survey` and click [OK]), select [Statistics], [Contingency tables], [Two-way table...], then select `Clap` as [Row variable] and `W.Hnd` as [Column variable], and click [OK]. Then the p-value will appear in the Output Window, though pairwise comparisons are not supported.

6 Relationship between the two quantitative variables

Two well-known methods to examine the relationship between the two quantitative variables are calculating correlation and fitting the regression models.

First of all, drawing scattergram is necessary. Let's consider the relationship between height and the span of spread writing hand in `survey` data set.

In Rgui console, type `plot(Wr.Hnd ~ Height, data=survey)`. If you would like to see the relationship separately for males and females, use `pch=as.integer(Sex)` option.

In Rcmdr, select [Graphs], [Scatterplot...], then select `Height` as [x-variable] and `Wr.Hnd` as [y-variable], check off the box beside "Smooth Line", and click [OK]. Plotting by different markers for males and females, click the [Plot by groups...] button and select `Sex` before the clicking final [OK].

6.1 The difference between correlation and regression

The correlation means the strength of the relationship between 2 variables, and the regression means how much the variance of a variable can be explained by the variance of the other variable, by fitting the linear model.

6.2 Correlation analysis

The relationship shown as scatterplot may be apparent or spurious correlation. The researcher must always pay attention to it.

To show the strength of correlation, Pearson's product moment correlation coefficients are usually used. As nonparametric (using rank) version, the Spearman's rank correlation coefficients are also used.

The definition of the Pearson's correlation coefficient r between the 2 variables X and Y is,

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

The null-hypothesis that the r is not different from 0 can be tested using t_0 value defined as follows and t-distribution with $n - 2$ degree of freedom.

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

In Rgui console, to calculate the Pearson's correlation coefficient between heights and spread writing hand spans with null-hypothesis testing, type as follows. For Spearman's rank correlation coeffients, `method="spearman"` option can be used.

```
cor.test(survey$Height, survey$Wr.Hnd)
```

In Rcmdr, select [Statistics], [Summaries], [Correlation test...], then select `Height` and `Wr.Hnd` (clicking with pressing Ctrl), and click [OK] (If you need Spearman's rank correlation, check the corresponding option). The following result will appear in the Output Window.

```
        Pearson's product-moment correlation

data:  survey$Height and survey$Wr.Hnd
t = 10.7923, df = 206, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5063486 0.6813271
sample estimates:
     cor
0.600991
```

The estimated r is 0.60 with 95% confidence interval of [0.50, 0.69]. The traditional criteria to judge the strength of correlation are, more than 0.7 'strong', 0.4-0.7 'moderate', 0.2-0.4 'weak'.

To calculate the correlations for males and females separately, we must make subset of the data. In Rgui console, it's easy to calculate those as follows.

```
males <- subset(survey, Sex=="Male")
cor.test(males$Height, males$Wr.Hnd)
females <- subset(survey, Sex=="Female")
cor.test(females$Height, females$Wr.Hnd)
```

In Rcmdr, select [Data], [Active data set], [Subset active data set], then type `Sex=="Male"` **in the box below "Subset expression" and type** `males` **in the box below "Name for new data set" and click [OK]. The active data set will automatically change from** `survey` **to** `males`**. Then, select [Statistics], [Summaries], [Correlation test...], and select** `Height` **and** `Wr.Hnd` **(clicking with pressing Ctrl), and click [OK]. You will find the estimate of correlation coefficient between** `Height` **and** `Wr.Hnd` **with the result of null-hypothesis testing of correlation coefficient being zero for males.**
To do similar calculation for females, at first you must change active dataset by clicking the box where `males` **is shown as active dataset, and select** `survey` **and click [OK]. Again, select [Data], [Active data set], [Subset active data set], then type** `Sex=="Female"` **in the box below "Subset expression" and type** `females` **in the box below "Name for new data set" and click [OK]. The active data set will automatically change from** `survey` **to** `females`**. Then, select [Statistics], [Summaries], [Correlation test...], and select** `Height` **and** `Wr.Hnd` **(clicking with pressing Ctrl), and click [OK]. You will find the estimate of correlation coefficient between** `Height` **and** `Wr.Hnd` **with the result of null-hypothesis testing of correlation coefficient being zero for females.**

6.3 Fitting a regression model

The principle of fitting regression models to observed data is that the variance of a dependent variable can be mostly explained by the variance of independent variables. If the explanatory power is enough, substituting the independent variables by actual values will serve a corresponding projection or estimation of the dependet variable. Reverse calculation is also possible, as in the case of so-called "working curve". A working curve (but often line, sometimes with transformation) provides the equation as regression model for the series of observed absorptions for fixed concentrations. Usually a working line can be used when its explanatory power is more than 98%.

If the zero-adjustment is done by standard solution with zero concentration, the regression line must go through the origin (therefore, intercept must be zero), otherwise (zero-adjustment is done by pure water) the regression line may not go through the origin.

For example, let the series of absorption for the standard solutions with fixed concentrations (0, 1, 2, 5, 10 $\mu g/\ell$) as (0.24, 0.33, 0.54, 0.83, 1.32), when the zero-adjustment was done by pure water. If we denote the absorption variable as y and the concentration variable as x, the regression model can be written as $y = bx + a$. The coefficients a and b (a is called as "intercept" and b is called as "regression coefficient") should be estimated by the least square method to find the set of a and b minimizing the sum of square errors,

$$f(a,b) = \sum_{i=1}^{5} (y_i - bx_i - a)^2$$

Solving the equations that each partial diffential of $f(a,b)$ by a and b equals 0, then we can obtain the following 2 equations.

$$b = \frac{\sum_{i=1}^{5} x_i y_i/5 - \sum_{i=1}^{5} x_i/5 \cdot \sum_{i=1}^{5} y_i/5}{\sum_{i=1}^{5} {x_i}^2/5 - \left(\sum_{i=1}^{5} x_i/5\right)^2}$$

$$a = \sum_{i=1}^{5} y_i/5 - b \cdot \sum_{i=1}^{5} x_i/5$$

Using these a and b values and measured absorption (for example 0.67), we can estimate the unknown concentration of sample solution. To note, the measured absorption of any sample must range within the values for standard solutions. The regression model has no guarantee to stand outrange of standard solutions[*22].

In Rgui console, we can apply `lm()` (linear model) to estimate the fitted regression model as follows.

```
y <- c(0.24, 0.33, 0.54, 0.83, 1.32)
x <- c(0, 1, 2, 5, 10)
# apply linear model fitting
res <- lm(y ~ x)
# show the summary of result
summary(res)
# draw scattergram with regression line
plot(y ~ x)
abline(res)
# calculate the concentration corresponding to the absorption of 0.67
(0.67 - res$coef[1])/res$coef[2]
```

The summary of result is shown below.

[*22] Such an extrapolation is not recommended. Usually concentrating or diluting the solutions to remeasure the absorption is recommended.

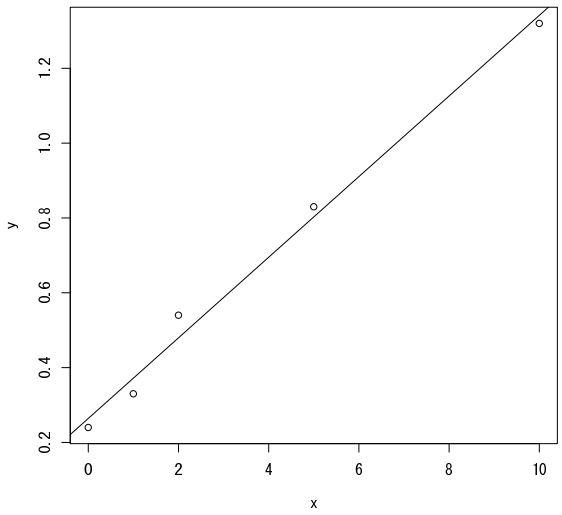
```
Call:
lm(formula = y ~ x)

Residuals:
       1        2        3        4        5
-0.02417 -0.04190  0.06037  0.02718 -0.02147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26417    0.03090   8.549 0.003363 **
x            0.10773    0.00606  17.776 0.000388 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04894 on 3 degrees of freedom
Multiple R-squared: 0.9906,     Adjusted R-squared: 0.9875
F-statistic:   316 on 1 and 3 DF,  p-value: 0.0003882
```

We can see that the estimated intercept was $a = 0.26417$ and regression coefficient was $b = 0.10773$, and the model explains 98.75% (0.9875) of the data, which is shown in `Adjusted R-squared`. The `p-value` means the result (significant probability) of testing the null-hypothesis (the extent how the variance of absorption can be explained by the model is similar to error variance).



The concentration for the absorption of 0.67 is given at last as 3.767084. Therefore, we can conclude that the concentration of the solution, of which absorption was 0.67, was 3.8 $\mu g/\ell$.

In Rcmdr, the data must be entered as a Data Set. Select [Data], [New data set] and type `workingcurve` in the box of "Enter name for data set:". After the "Data Editor" window appears, click [var1] and type `y` in the "variable name" box and check the radio button of "numeric" as type and press Enter key. Next, change [var2] to [x] similarly. Then enter the data into each cells, and close the window (usually select [File], [Close]).
To draw scattergram with regression line, select [Graphs], [Scatterplot...], then select `x` as "x-variable" and `y` as "y-variable". Check off the box beside "Smooth Line" and click [OK].
To apply a linear regression model fitting, select [Statistics], [Fit models], [Linear regression], then select `y` as "response variable" and `x` as "Explanatory variables". Clicking [OK] leads to the summary of results shown in the Output Window.

For other situations than working curves, linear regression models can be applied in a similar manner. Let's go back to the example of `survey` data set[*23]. If we want to explain the variance of the span of writing hand by the height, we can apply the linear regression model to the `survey` data set by typing as follows in Rgui console.

```
res <- lm(Wr.Hnd ~ Height, data=survey)
summary(res)
```

In Rcmdr, after activating `survey` as already mentioned, select [Statistics], [Fit models], [Linear regression], then select `Wr.Hnd` as "response variable" and `Height` as "Explanatory variables". Clicking [OK] gives the summary result.

6.4 Testing the stability of estimated coefficients

When the response variable has virtually no relationship with the explanatory variable, the sums of squared residuals for many possible regression lines (any line on the centroid) may give almost same values. In other words, the estimated intercept and slope are very unstable in such situation. To evaluate the stability of parameters of regression line (regression coefficient b and intercept a), t_0 values are usually used. Let the relationship between y and x be expressed by the equation of $y = a_0 + b_0 x + e$, and assume the error term e obeying the normal distribution with mean 0 and variance σ^2, the estimated regression coefficient a would obey the normal distribution of mean a_0, variance $(\sigma^2/n)(1 + M^2/V)$, where M and V are the mean and the variance of x. Then the sum of squared residuals Q divided by the variance of error σ^2 (say, Q/σ^2) obeys the chi-square distribution with degree of freedom $(n-2)$. Therefore, the $t_0(a_0)$ defined as follows obeys the t-distribution with the degree of freedom $(n-2)$.

$$t_0(a_0) = \frac{\sqrt{n(n-2)}(a - a_0)}{\sqrt{(1 + M^2/V)Q}}$$

However, to calculate this value, a_0 must be known. Under the null hypothesis of $a_0 = 0$, $t_0(0)$ calculated from the observed data is almost matching with $t_0(a_0)$ and obeys the t-distribution with degree of freedom $(n-2)$. The absolute value of $t_0(0)$ is less than the 97.5% point of t-distribution at the 95% probability. We can also get the significance probability using the distribution function (cumulative probability density function).

Similarly, we can calculate $t_0(b)$ for regression coefficient as follows.

$$t_0(b) = \frac{\sqrt{n(n-2)V}b}{\sqrt{Q}}$$

Using the relationship that the $t_0(b)$ obeys the t-distribution with degree of freedom $(n-2)$, we can calculate the significance probability.

[*23] Of course, the `MASS` package must be loaded before using `survey` data set.

If the significant probability is very small (usually less than 5%, this criteria is called as the significance level of the test), we can say that a_0 or b_0 is significantly different from zero, which means the stability of estimated a_0 or b_0.

In both Rgui concole and Rcmdr, the significance probabilities are shown at the column of `Pr(>|t|)`.

7 Applied regression models

7.1 Multiple regression model

The explanatory variables can include two or more variables. In such case, the model is called as "multiple regression model". There are some points to pay attention, but basically the explanatory variables can be given as the right terms of linear model, connected by +. For example, for the same data previously described, if you would like to explain the variance of the span of writing hand (`Wr.Hnd`) by the variance of height (`Height`) and the variance of the span of non-writing hand (`NW.Hnd`), you may type as follows in Rgui console.

```
res <- lm(Wr.Hnd ~ Height + NW.Hnd, data=survey)
summary(res)
```

In Rcmdr, select [Statistics], [Fit models], [Linear regression], then select `Wr.Hnd` as "response variable", and `Height` and `NW.Hnd` with pressing Ctrl key as "Explanatory variables". Clicking [OK] leads to the summary of results shown in the Output Window.

In the multiple regression model, the estimated regression coefficients are the "partial regression coefficients", which adjust the effects of other explanatory variables on the response variable to obtain each explanatory variable's own effect on the response variable. But the values of partial regression coefficients depend on the absolute scale of each variable, so that those cannot show the relative strength of effect on the response variable for each explanatory variable. For such comparison, the standardized partial regression coefficients can be used. In Rgui console, type as follows, then you obtain the estimates as `stb` for the standardized partial regression coefficients.

```
sdd <- c(0,sd(res$model$Height),sd(res$model$NW.Hnd))
stb <- coef(res)*sdd/sd(res$model$Wr.Hnd)
stb
```

The Rcmdr does not provide this as a menu item, but you can do so by editing the commands in script window, selecting lines and click [Submit].

7.2 Evaluation of the goodness of fit

It is always necessary to evaluate the goodness of fit of the regression model to the data.

After the least square estimation of a and b, we can define $z_i = a + bx_i$ for each x. Then $e_i = y_i - z_i$ can be considered as "residuals". The residuals is the remaining part of the variance of y_i, that could not be explained by the regression model. Thus, the greater the residuals are, the worse the goodness of fit is. We would like to treat the both plus and minus residuals in its absolute distance, so that we can define the sum of squared residuals, Q, as follows.

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - z_i)^2$$

$$= \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n - \frac{(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}/n$$

The sum of squared residuals Q is the scale to show badness of fit of the regression model. Q divided by n is called as residual variance (we will denote it as var(e)).

Among the values of var(e) and the variance of y (var(y)) and Pearson's product limit correlation coefficient r, the following equation always stands.

$$\text{var}(e) = \text{var}(y)(1 - r^2)$$

Therefore,

$$r^2 = 1 - \frac{\text{var}(e)}{\text{var}(y)}$$

Thus the closer to 1 r^2 is, the goodness of fit is higher. In that meaning, r^2 is called as "deterministic coefficient" or "the attributable proportion of x".

Since r^2 becomes larger depending on the number of explanatory variables, usually the r^2 will be adjusted for the degree of freedom. That is, `Adjusted R-Squared` in the summary of results.

As another indicator for the goodness of fit, the AIC (Akaike information criterion) is sometimes used (particularly in multiple regression models), which can be calculated in R, using the resulted object of linear model fitting (for example, `res` of the example above). In Rgui console, type `AIC(res)`, then you can get the AIC value. Here I don't explain any more, but many online materials and books can be found[*24].

7.3 Points to be paid attention in fitting regression model

The target variables may be measurements including error. In such situation, it is not valid to assume one as response variable and the other as explanatory variable. Generally speaking, if we can assume the direction of effect like the stature determining weight and not *vise versa*, the regression is possible where the stature is explanatory variable and the weight is response variable. However, when the explanatory variable includes measurement error, the explanatory power of the regression model become worse. In addition, the regression models with opposite combination of response variable and explanatory variable do not match. Thus, it is very important that the determination of which variable is response variable shoule be based on the direction of causal relationship, with enough reference to previous studies and clinical/biological knowledge.

Another point to be noticed is extrapolation of regression model for prediction. Especially the extrapolation should be avoided when you apply the working curve for prediction, because the linearity of the working curve is only confirmed within the range of standard material concentrations. The increase of absorbance tends to be smaller in higher concentration ranges due to saturation of molecules, the linearity is lost there. If you measure the samples with high concentration, you must dilute them into the ranges of standard materials.

Exercise

A built-in dataset `airquality` includes the air quaility data in New York from May to September 1973. The variables are `Ozone` for ozone gas concentration in ppb, `Solar.R` for solar radiation in lang, `Wind` for wind speed in mph, `Temp` for atomospheric temperature in degree F, `Month` in number (5-9) and `Day` in number (1-31).
Let's fit the regression model for this data with ozone gas concentration as response variable and solar radiation as explanatory variable.

In Rgui console, enter the following 4 lines.

```
plot(Ozone ~ Solar.R, data=airquality)
```

[*24] `http://en.wikipedia.org/wiki/Akaike_information_criterion` is the explanation in the Wikipedia.

```
res <- lm(Ozone ~ Solar.R, data=airquality)
abline(res)
summary(res)
```

In Rcmdr, at first, the `airquality` **must be activated by select [Data], [Data in packages], [Read data set from an attached packages ...], then double-click** `datasets` **in the left panel and double-click** `airquality` **in the right panel, then click [OK].**
To draw scattergram, [Graphs], [Scatterplot ...], then select `Solar.R` **as x-variable and** `Ozone` **as y-variable. Check the box beside "Smooth Line" off, and click [OK]. Then you will get the scattergram with a regression line. To obtain the numerical result of regression model fitting, select [Statistics], [Fit models ...], [Linear regression], then select** `Ozone` **as Response variable and** `Solar.R` **as Explanatory variables, and click [OK]. Then you will see the result in the Output window.**

Both give the same results as follows.

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

Residuals:
    Min      1Q  Median      3Q     Max
-48.292 -21.361  -8.864  16.373 119.136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.59873    6.74790   2.756 0.006856 **
Solar.R      0.12717    0.03278   3.880 0.000179 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 31.33 on 109 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-Squared: 0.1213,Adjusted R-squared: 0.1133
F-statistic: 15.05 on 1 and 109 DF,  p-value: 0.0001793
```

The fitted regression model is Ozone $= 18.599 + 0.127 \cdot$ Solar.R, and the `p-value` shown in the bottom line is 0.0001793, as the result of F-test. Therefore, the fitting of the model can be judged as significant. However, the `Adjusted R-squared` shown above line is 0.11, which means only about 10% of the variance of the Ozone concentration can be explained by this model. We should judge the model is not so good.

To improve fitting, it may be possible to add more variables (for example, `Wind` and/or `Temp`) as explanatory variables as the multiple regression model. In Rgui console, you can easily do so by typing the next 3 lines. Then you see about 60% of `Adjusted R-Squared` value.

```
mres <- lm(Ozone ~ Solar.R + Wind + Temp, data=airquality)
summary(mres)
AIC(mres)
```

In Rcmdr, select [Statistics], [Fit models ...], [Linear model], then type `Ozone` **at the box to the left of ~ and type** `Solar.R + Wind + Temp` **at the box to the right of ~, and click [OK].**

7.4 Analysis of Covariance (ANACOVA/ANCOVA)

When the same response variable and explanatory variables are investigated for some groups, we often want to check the differences of the regression coefficients by the groups. In such case, the analysis of covariance is applicable.

Typical model is,

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{C} + \beta_2 \mathrm{X} + \beta_{12} \mathrm{C} \cdot \mathrm{X} + \varepsilon$$

where **C** is binary variable and **X** and **Y** are numerical (continuous) variables. We would like to compare the means of **Y** between the 2 categories of **C**, but considering the significant correlation between **Y** and **X**, to compare the adjusted means of **Y** between the 2 categories of **C**, each adjusted for **X**, only when the slopes of linear regression of **X** on **Y** are not different. The adjusted mean can be calculated as the sum of the coefficient of **C** and the mean of covariate **X** times its coefficient and the intercept.

As prerequisites, the regression models for both categories of **C** show good fits for the data. If the models poorly fit the data, stratified analysis for each category of **C** may be recommended. The typical procedure is described below.

(1) Testing the null hypothesis of equal slopes Test $H_0 : \beta_1 = \beta_2$ and $H_1 : \beta_1 \neq \beta_2$. For each category group, calculate the mean, variation, and covariation[*25], the mean square error d_1 under the alternative hypothesis H_1 can be calculated as,

$$d_1 = SS_{Y1} - (SS_{XY1})^2/SS_{X1} + SS_{Y2} - (SS_{XY2})^2/SS_{X2}$$

and the mean square error d_2 under the null hypothesis H_0 can be calculated as,

$$d_2 = SS_{Y1} + SS_{Y2} - (SS_{XY1} + SS_{XY2})^2/(SS_{X1} + SS_{X2})$$

then we can obtain $F = (d_2 - d_1)/(d_1/(N-4))$, which obeys F-distribution with the 1st d.f. 1 and the second d.f. $N-4$ under the null hypothesis H_0.

(2) If equal slopes, testing the null hypothesis of equal intercept Based on $\beta_1 = \beta_2$, estimate the common slope β as,

$$\beta = (SS_{XY1} + SS_{XY2})/(SS_{X1} + SS_{X2})$$

then test $H_0' : \alpha_1 = \alpha_2$ and $H_1' : \alpha_1 \neq \alpha_2$. Under the null hypothesis H_0', calculate the mean square error $d_3 = SS_Y - (SS_{XY})^2/SS_X$, then get $F = (d_3 - d_2)/(d_2/(N-3))$, which should obey the F-distribution with 1st d.f. 1 and 2nd d.f. $N-3$. When the H_0' is rejected, substitute the mean of each group by the actual value using the common slope, then we can get the intercept of each group. If the H_0' is not rejected, usual linear regression should be applied to the pooled data.

(3) If significantly different slopes, stratified analysis Separately estimate $\beta_1 = SS_{XY1}/SS_{X1}$ and $\beta_2 = SS_{XY2}/SS_{X2}$, then α_1 and α_2 should also be estimated by each linear regression equation, with substituting the explanatory variable by its actual mean.

Exercise

`http://phi.med.gunma-u.ac.jp/grad/sample3.dat` is a text datafile delimited with tab, which includes variables for the name of prefecture (`PREF`), the region whether `East` or `West` (`REGION`), the number of cars per 100 households in 1990 (`CAR1990`), the number of deaths by traffic accident per 100,000 population in 1989 (`TA1989`), the proportion (in %) of dwellers in densely inhabited area according to 1985 national census (`DIDP1985`).

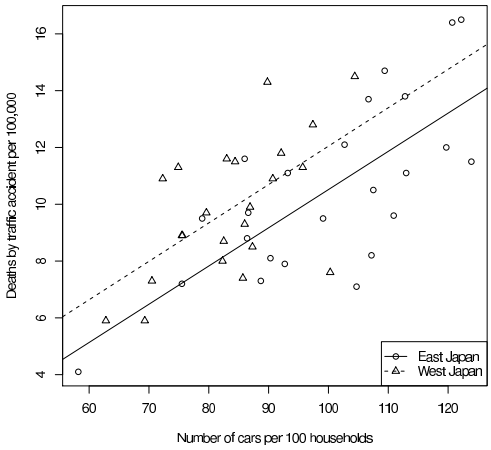
For this data, examine the difference between `East` and `West` of `TA1989`, adjusting the effect of `CAR1990`, using ANACOVA[a].

[a] Note: There is no significant difference of the number of deaths by traffic accident between East and West Japan, without adjustment for car holding.

[*25] For the first group with sample size $N1$, denote the ith elements as x_i and y_i, then $E_{X1} = \sum x_i/N1$, $SS_{X1} = \sum(x_i - E_{X1})^2$, $E_{Y1} = \sum y_i/N1$, $SS_{Y1} = \sum(y_i - E_{Y1})^2$, $E_{XY1} = \sum x_i y_i/N1$, $SS_{XY1} = \sum(x_i y_i - E_{XY1})^2$. For the second group, the mean, variation and covariation can be caluculated similarly.

In Rgui console, type as follows.

```
sample3 <- read.delim("http://phi.med.gunma-u.ac.jp/grad/sample3.dat")
plot(TA1989 ~ CAR1990, pch=as.integer(REGION), data=sample3,
 xlab="Number of cars per 100 households", ylab="Deaths by traffic accident per 100,000")
east <- subset(sample3,REGION=="East")
regeast <- lm(TA1989 ~ CAR1990, data=east)
summary(regeast)
west <- subset(sample3,REGION=="West")
regwest <- lm(TA1989 ~ CAR1990, data=west)
summary(regwest)
abline(regeast,lty=1)
abline(regwest,lty=2)
legend("bottomright",pch=1:2,lty=1:2,legend=c("East Japan","West Japan"))
summary(lm(TA1989 ~ REGION*CAR1990, data=sample3))
anacova <- lm(TA1989 ~ REGION+CAR1990, data=sample3)
summary(anacova)
cfs <- dummy.coef(anacova)
cfs[[1]] + cfs$CAR1990 * mean(sample3$CAR1990) + cfs$REGION
```



The p-value for the null hypothesis that the coefficient of REGION in the last model is not different from zero was 0.0319. There proved to be a significant difference of TA1989, adjusted for CAR1990, between East and West. Adjusted means for regions is obtained as follows.

```
    East     West
 9.44460 10.96650
```

In Rcmdr, at first, the `sample3` must be read from internet, select [Data], [Import Data], [from text file, clipboard, or URL...], then type `sample3` in the box of "Enter name for data set:", check the ratio button beside "Internet URL", and check the radio button beside [Tabs] of "Field Separator" and click [OK], then type `http://phi.med.gunma-u.ac.jp/grad/sample3.dat` as Internet URL and click [OK].
To draw scattergram, [Graphs], [Scatterplot ...], then select `CAR1990` as x-variable and `TA1989` as y-variable. Check the box beside "Smooth Line" off, and click the button of "Plot by group" and select `REGION`. Clicking `OK`, then you will get the scattergram with two regression lines. To test the difference of slopes of the two regression lines, select [Statistics], [Fit models ...], [Linear model], then select `TA1989` as Response variable and `REGION+CAR1990+REGION:CAR1990` as Explanatory variables, and click [OK]. Then you will see the result in the Output window. The p-value in the line of REGIONWest:CAR1990 is 0.990, which means no difference of the slopes in East and West. Then to test the difference of the mean TA1989 adjusted for CAR1990, select [Statistics], [Fit models ...], [Linear model], then select `TA1989` as Response variable and `REGION+CAR1990` as Explanatory variables, and click [OK]. The p-value in the line of REGIONWest is 0.0319, which means statistically significant difference between the adjusted means. But adjuted means itself can only be obtained by the commands written above.

7.5 Logistic regression analysis

The logistic regression is a kind of fitting the regression model to the binary data. The reponse variable is not continuous but binary, and obeys the binary distribution. Therefore in Rgui console, we use `glm()` instead of `lm()`.

The logistic regression model fitting is often used in medical statistics. For example, let the response variable whether having disease or not and let the explanatory variables whether having risk factors or not with some confoundings like age, applying the logistic regression model will lead to calculate the odds ratio for each risk factor with controlling the effects of other risk factors and confoundings simultaneously.

Usually whether having disease or not is expressed as 1 or 0, as binary variable. The data mean the prevalence proportion, how much proportion of the total have disease. Therefore, the left term of the logistic regression equation ranges within 0 and 1, but the right terms of the equation are composed of several categorical variables with (sometimes numeric) confoundings, and thus ranges over all real values from minus infinity to infinity. Therefore, in the logistic regression, the left term is transformed by the logit transformation (taking natural logarithm of the ratio of itself to 1 minus itself).

Thus, let the prevalence proportion of the disease P, the logistic regression model is,

$$\ln(P/(1-P)) = b_0 + b_1 X_1 + ...b_k X_k$$

If X_1 is the binary variable meaning whether having a risk factor or not, and $X_2, ...X_k$ are confoundings, subtracting the case of $X_1 = 0$ from the case of $X_1 = 1$,

$$b1 = \ln(P_1/(1-P_1)) - \ln(P_0/(1-P_0)) = \ln(P_1 * (1-P_0)/(P_0 * (1-P_1)))$$

Thus, b_1 means the logarithm of the odds ratio controlling all other variables. Assuming the log-odds ratio to obey the normal distribution, the 95% confidence intervals of the odds ratio can be obtained by the following formula.

$$\exp(b_1 \pm 1.96 \times \mathrm{SE}(b_1))$$

Exercise

The builit-in dataset `birthwt` included in the `library(MASS)` is the record of 189 births at Baystate Medical Center in Springfield, for the relationship between low birth weight and its risk factors. Included variables are listed below. Apply the logistic regression analysis where the response variable is `low`, which means whether the baby's weight at birth is less than 2.5 kg or not.

low	A binary variable for low birth weight (1, when birth weight is less than 2.5 kg, 0 otherwise)
age	Mother's age at birth in years
lwt	Mother's weight (in pounds[a]) at last menstrual period
race	Mother's race (1 = white, 2 = black, 3 = other)
smoke	Smoking status during pregnancy (1, when smoked)
ptl	Number of previous premature labours (Note: This means preterm births!!)
ht	History of hypertension (1, when experienced)
ui	Presence of uterine irritability (1, present)
ftv	Number of physician visits during the first trimester
bwt	Baby's birth weight in grams

[a] Abbreviated as lbs, where 1 lb. = 0.454 kg.

The data include many variables, and the logistic regression analysis should use the target risk factor variables with all possible confouding factors as explanatory variables. Any variable having correlation with both of response variable and explanatory variable may be a confounding factor.

Here the response variable is `low` and we assume that, based on the detailed analysis and consideration, explanatory variables should be `race`, `smoke`, `ht`, `ui`, `lwt`, and `ptl`. Before using them as explanatory variables, we must **transform** the type of variable, except for truely numeric `lwt` and `ptl`, **from numeric into factor**. The response variable also must be transformed from numeric into factor.

In Rgui console, we can do so by typing as follows.

```
library(MASS)
data(birthwt)
birthwt$clow <- factor(birthwt$low, labels=c("NBW","LBW"))
birthwt$crace <- factor(birthwt$race, labels=c("white","black","others"))
birthwt$csmoke <- factor(birthwt$smoke, labels=c("nonsmoke","smoke"))
birthwt$cht <- factor(birthwt$ht, labels=c("normotensive","hypertensive"))
birthwt$cui <- factor(birthwt$ui, labels=c("uterine.OK","uterine.irrit"))
```

In Rcmdr, at first select [Tools], [Load package(s)...], then select `MASS` and click [OK] button. Next select [Data], [Data in packages], [Read data set from an attached package...], then select (double-click) `MASS` in the left panel and select (double-click) `birthwt` in the right panel, and click [OK].
Then select [Data], [Manage variables in active data set], [Convert numeric variables to factors...], then select `ht` and type `cht` into the box besides "New variable name or prefix for multiple variables:", then type corresponding level names such as `normotensive` for level 0 and `hypertensive` for level 1. For `low`, `race`, `smoke`, and `ui`, do the conversion in a similar manner.

Fitting the logistic regression model to this data can be done by typing as follows in Rgui console.

```
res <- glm(clow ~ crace+csmoke+cht+cui+lwt+ptl, family=binomial(logit), data=birthwt)
summary(res)
```

If you evaluate how the model explains the data, Nagelkerke's R-square can be used, instead of adjusted R-squares in the case of multiple linear regression model, as follows.

```
require(fmsb)
```

```
NagelkerkeR2(res)
```

To conduct the logistic regression analysis by Rcmdr, select [Statistics], [Fit models], [Generalizd linear models...], then put the cursor at the left box of ~ and click `clow` or type `clow` there, and type at the right box of ~ as `crace+csmoke+cht+cui+lwt+ptl` (or click variables and mathematical symbols). Then select `binomial` at the box of "Family (double-click to select)" and `logit` at the box of "Link function", and click [OK]. The Nagelkerke's R square is not supported in Rcmdr.

Both Rgui console and Rcmdr generate the same result as shown in the box below.

```
Call:
glm(formula = clow ~ crace + csmoke + cht + cui + lwt + ptl,
    family = binomial(logit), data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9049  -0.8124  -0.5241   0.9483   2.1812

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.086550   0.951760  -0.091  0.92754
craceblack        1.325719   0.522243   2.539  0.01113 *
craceothers       0.897078   0.433881   2.068  0.03868 *
csmokesmoke       0.938727   0.398717   2.354  0.01855 *
chthypertensive   1.855042   0.695118   2.669  0.00762 **
cuiuterine.irrit  0.785698   0.456441   1.721  0.08519 .
lwt              -0.015905   0.006855  -2.320  0.02033 *
ptl               0.503215   0.341231   1.475  0.14029
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 201.99  on 181  degrees of freedom
AIC: 217.99

Number of Fisher Scoring iterations: 4
```

From this output, the result is usually summarized as the table shown below. Like this, continuous variables are usually described as "included in the model as covariates", but sometimes included in the table with odds ratio. The estimated coefficients in the output are log odds ratios, so that it is usually transformed to odds ratios by taking exponential with 95% confidence intervals. This procedure is not supported in Rcmdr, so you need to type the followings in the Rgui console, if the name of the model is `GLM.1`.

```
exp(coef(GLM.1))
exp(confint(GLM.1))
```

Table. The result of logistic regression analysis for the risk factors of low birth weight babys at Baystate Medical Center.

		95% confidence intervals		
Explanatory variables*	Odds Ratio	Lower limit	Upper limit	p-value
Race (White)				
Black	3.765	1.355	10.68	0.011
Other colored	2.452	1.062	5.878	0.039
Smoke (No smoke)	2.557	1.185	5.710	0.019
Hypertensive (Normotensive)	6.392	1.693	27.3	0.008
Uterine irritability (Normal uterine)	2.194	0.888	5.388	0.085

AIC: 217.99, D_{null}: 234.67 (d.f. 188), D: 201.99 (d.f. 181)

* Referene category is shown in parenthesis. The logistic regression model includes mother's weight at the last menstruation and number of previous preterm births as covariates.

8 Contingency tables for independence hypothesis

How can we examine the relationship between categorical variables? Of course, the logistic regression model can also be applied to the relationship between the 2 categorical variables. In addition, as well as Pearson's correlation coefficient for 2 numeric variables, there are many indices to show the strength of relationship for 2 categorical variables: For example, the phi coefficient (usually written as ρ) is the same calculation with the Pearson's correlation coefficient where the values are expressed as 0 or 1 (0 means "not having" and 1 means "having" for each of cause and disease). Denote the proportion of having cause among patients as θ_1, and the proportion of having cause among healthy controls as θ_2, then

$$\rho = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$$

As another way to show the strength of relationships, in many epidemiologic studies, taking the ratios or differences between the incidence rates, risks or odds for the exposed (or having causes) group and nonexposed group. The more apart from 1 the ratio is or the more apart from 0 the difference is, the strength of relationship is stronger. For example, incidence rate ratio and odds ratio (the same mean as in the case of logistic regression, but here ignoring the effect of other variable) are frequently calculated in epidemiologic studies. In this practical lesson, we do not treat these ratios or differences for the shortness of time. I recommend you to read the textbook of epidemiology written by Kenneth J. Rothman "Epidemiology: An Introduction", Oxford Univ. Press, 2002. And, such epidemiologic analysis can be conducted using the additional packages like `epitools`, `vcd`, and `Epi` in CRAN.

Here we do the test for independence of 2 categorical variables, where the null hypothesis is that those 2 categorical variables are independent.

Statistical information of categorical variable is the frequencies of each category. Therefore, the relationship between the 2 categorical variables can be checked by making a 2 dimentional contingency table. In Rgui console, `table()` function can be used to make contingency table. Usually it is called as cross table*[26]. As we have already seen, the test of proportions by normal approximation is mathematically equivalent to the test of chi-square test of contingency table, but the detailed explanation will be given below.

8.1 Chi-square test for independence

Concerning the test for independence, chi-suqare test is the most popular. That is, a kind of "goodness of fit" test. The null hypothesis is that the two categorical variables has no correlation (in other words, the two variables are independent). Actually, the procedure is (1) calculate expected numbers of each combination if 2 variables are independent, (2) for each combination, calculate the difference of obserbed number and expected number, square it, divide it by the expected number,

*[26] Especially the two categorical variables are both binary variable, the table is called as "2 by 2 cross tabulation" (2 by 2 contingency table), and its statistical property is well described in many textbook.

(3) sum up them to calculate the chi-square statistic, which obeys chi-square distribution with d.f. 1, (4) if the chi-square statistic is greater than 3.84 (95% point of chi-square distribution with d.f. 1), we can conclude that the difference between observation and expectation under independent hypothesis is significantly larger than zero, thus the null hypothesis can be rejected at 5% significance level.

	A	$\bar{A}$
B	a	b
$\bar{B}$	c	d

If the values which can be taken by 2 categorical variable A and B are limited to "having" and "not having", the combination of these 2 variables are only 4 cases as "having both A and B ($A \cap B$)", "not having A but having B", "having A but not having B" and "not having A nor B". When the numbers of all those combinations are summarized as the table shown above, the probability structure of the population can be written as the next table.

	A	$\bar{A}$
B	π_{11}	π_{12}
$\bar{B}$	π_{21}	π_{22}

If this table is given, expected numbers of all combinations can be summarized as the next table, where $N = a + b + c + d$.

	A	$\bar{A}$
B	$N\pi_{11}$	$N\pi_{12}$
$\bar{B}$	$N\pi_{21}$	$N\pi_{22}$

From the two tables (observed and expexted numbers of all combinations), χ^2 statistic can be calculated by the formula:

$$\chi^2 = \frac{(a - N\pi_{11})^2}{N\pi_{11}} + \frac{(b - N\pi_{12})^2}{N\pi_{12}} + \frac{(c - N\pi_{21})^2}{N\pi_{21}} + \frac{(d - N\pi_{22})^2}{N\pi_{22}}$$

Then the χ^2 statistic can be statistically tested using chi-square distribution with d.f. 3. However, usually $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$ are unknown.

Assuming that $Pr(\bar{A}) = 1 - Pr(A)$ and $Pr(A \cap B) = Pr(A)Pr(B)$ under the null hypothesis[*27], we can estimate 2 parameters $Pr(A)$ and $Pr(B)$[*28].

The point estimate of $Pr(A)$ is naturally proved to be $(a+c)/N$, because it can be considered as the proportion of "having" A, ignoring B. Similarly, the point estimate of $Pr(B)$ is of course $(a + b)/N$. Based on these, πs can be obtained as follows.

$$\pi_{11} = Pr(A \cap B) = Pr(A)Pr(B) = (a + c)(a + b)/(N^2)$$

$$\pi_{12} = (b + d)(a + b)/(N^2)$$

$$\pi_{21} = (a + c)(c + d)/(N^2)$$

$$\pi_{22} = (b + d)(c + d)/(N^2)$$

Using these values, the χ^2 statistic can be obtained from the following equation.

$$\chi^2 = \frac{\{a - (a + c)(a + b)/N\}^2}{\{(a + c)(a + b)/N\}} + \frac{\{b - (b + d)(a + b)/N\}^2}{\{(b + d)(a + b)/N\}} + \frac{\{c - (a + c)(c + d)/N\}^2}{\{(a + c)(c + d)/N\}} + \frac{\{d - (b + d)(c + d)/N\}^2}{\{(b + d)(c + d)/N\}}$$

[*27] This null hypothesis means that the numbers of individuals in each combination are proportionate to the distribution of sums of "having" A and "not having" A separately for "having" B and "not having" B. Therefore, this test is mathematically equivalent with the test of the differences of proportions where the null hypothesis is that the "having" A proportion in "having" B equals to that in "not having" B.

[*28] Here $Pr(X)$ denotes the probability of "having" X. In addition, here we estimate the 2 parameters $Pr(A)$ and $Pr(B)$ from data, the degree of freedom for the chi-square distribution in which the obtained chi-square statistic obeys to is 1, which is 3 − 2.

$$= \frac{(ad-bc)^2\{(b+d)(c+d)+(a+c)(c+d)+(b+d)(a+b)+(a+c)(a+b)\}}{(a+c)(b+d)(a+b)(c+d)N}$$

The in-between terms of { and } in the numerator is N^2, so that the above equation can be simplified as the following equation.

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

However, this equation is usually slightly modified by Yates' continuity correction. The reason of this correction is that the approximation of chi-square distribution can be improved by adding or subtracting 0.5 from the actual frequencies of each combination. Then the following equation is obtained.

$$\chi_c{}^2 = \frac{N(|ad-bc|-N/2)^2}{(a+c)(b+d)(a+b)(c+d)}$$

The $\chi_c{}^2$ obeys the chi-square distribution with d.f. 1. To note, when $|ad-bc|$ is less than $N/2$, Yates' continuity correction doesn't make sense, so that usually let χ^2 zero. Nonetheless, R's `chisq.test()` function applys the correction even if $|ad-bc| < N/2$, because the R development core team take the position of Yates' original article. FYI, `prop.test()` doesn't take this position and $\chi^2 = 0$ when $|ad-bc| < N/2$.

Cross-tabulation in Rgui console can be done by `table()` or `xtabs()` function. The resulted object is a matrix object with `table` class. If the resulted matrix is known, it can be directly given as a matrix object. For example, when a=12, b=8, c=9, and d=10, you can define the matrix object `x` and conduct chi-square test as follows.

```
x <- matrix(c(12,9,8,10), 2, 2)
# x <- matrix(c(12,8,9,10), 2, 2, byrow=TRUE) is also possible.
chisq.test(x)
```

In Rcmdr, select [Statistics], [Contingency tables], [Enter and analyze two-way table ...], then directly enter the frequencies into the corresponding cells and click [OK]. The result of chi-square test will be shown in the Output Window.

Exercise

For 100 lung cancer patients, select 100 healthy controls with same age and sex one by one (such controls are called as pair-match sampling[a].
For each group, the result of asking smoking habits, 80 among 100 patients and 55 among 100 controls had experience of smoking.
Test the null-hypothesis that lung cancer has no relationship with smoking, using chi-square test.

[a] Caution: Use of pair-match sampling in such case-control study may impair the representativeness of controls from general population.

In other words, the null-hypothesis is that suffering from lung cancer is independent from past smoking habits. The 2 by 2 contingency table is shown below.

	Lung cancer patients	Healthy controls	Total
Past smoker	80	55	135
Never smoked	20	45	65
Total	100	100	200

Under the null-hypothesis that lung cancer is not related with smoking, expected numbers of each combinations can be calculated as follows.

	With lung cancer	Without lung cancer
Smoking	$135 \times 100/200 = 67.5$	$135 \times 100/200 = 67.5$
Nonsmoking	$65 \times 100/200 = 32.5$	$65 \times 100/200 = 32.5$

Therefore, the chi-square value with Yates' continuity correction is obtained as follows.

$$\chi_c^2 = (80-68)^2/67.5 + (55-67)^2/67.5 + (20-32)^2/32.5 + (45-33)^2/32.5 = 13.128...$$

Type `1-pchisq(13.128,1)`, then `[1] 0.00029...` is shown as the significant probability of chi-square test. Thus the null-hypothesis is rejected at the significance level of 5%. Consequently the association between lung cancer and smoking was statistically significant.

In Rgui console, the above result can be easily obtained by typing the one line below.

```
chisq.test(matrix(c(80,20,55,45),2,2))
```

In Rcmdr, select [Statistics], [Contingency tables], [Enter and analyze two-way table ...], then directly enter the frequencies into the corresponding cells and click [OK].

Exercise

Test the null-hypothesis that sex (`Sex`) is independent from writing hand (`W.Hnd`) for the `survey` data in `MASS` library.

In Rgui console, type as follows.

```
require(MASS)
chisq.test(xtabs(~ Sex+W.Hnd, data=survey))
```

The `p-value` will be obtained as 0.6274, which means no significant association existing between sex and writing hand.

In Rcmdr, after activating `survey` data set, select [Statistics], [Contingency tables], [Two-way table ...], then select `Sex` as "Row variable" and `W.Hnd` as "Column variable", and click [OK]. The result shown in the Output Window as `X-squared = 0.5435, df = 1, p-value = 0.461`, which is without Yates' continuity correction.
In Rcmdr, there is no way to use `chisq.test()` function with the option `correct=TRUE`.
However, it is possible to get the same result with `chisq.test()` with Yates' continuity correction in a different way. Select [Statistics], [Proportions], [Two-sample proportions test ...], then select `Sex` as "Groups" and `W.Hnd` as "Response Variable", and click the radio button beside "Normal approximation with continuity correction", then click [OK]. In this way, you can get 0.6274 as `p-value`.

8.2 Fisher's exact probability

When you find very low expected number of any combination, normal approximation in chi-square test is very bad. In such case, re-categorizing may decrease types of combinations and increase the expected numbers of each combination, but much better solution is provided as Fisher's exact probability (test).

For a given two-way table, if we can fix the marginal frequencies of the two-way tables (in other words, the proportions of each category are fixed for each variable), it is possible to calculate the probabilities (obeying hypergeometric distribution) of eventually obtaining the combination for all possible tables. Then, the sum of the probabilities equal to or lower than the given table's probability means the eventually obtaining such table under the null-hypothesis that there is no relationship between the two variables. The probability calculated in this manner is called as Fisher's exact probability. This is not approximation, so that the small number of samples and very low expected number of combinations are not the problem.

The method can be formulated as follows[*29]. Let's assume the finite population with size N, whose data of two variables `A` and `B` are given. Among N individuals, let's denote the number of the individuals whose data of variable `A` is 1 as m_1 and $m_2 = N - m_1$. In the situation that n_1 is the number of the individuals whose data of variable `B` is 1 (of course, $n_2 = N - n_1$),

[*29] Here the explanation is given for 2 by 2 table, but it is applicable for any two-way table.

and let a be the number of individuals whose data for both variable `A` and `B` are 1. We can calculate the probability that a exactly equals the number of the individuals whose data of variable `A` is 1 among n_1.

This probability p_a is the products of the number of combinations to extract a from m_1 and the number of combinations to extract $n_1 - a$ from m_2, divided by the number of combinations to extract n_1 from N. Therefore, the probability that the null-hypothesis "`A` and `B` are independent" can stand is the summation of all the probabilities that are equal to or smaller than p_a among all possible tables.

The extracting process is sampling from finite population without replacement, so that the mean $E(a)$ and the variance $V(a)$ can be given by the following equations.

$$E(a) = n_1 m_1 / N$$

$$V(a) = \{(N - n_1)/(N - 1)\} n_1 (m_1/N)(m_2/N) = (m_1 m_2 n_1 n_2)/\{N^2(N - 1)\}$$

Actually this calculation requires so enourmous computation that the statistical software will do it. To note, the probabilities for possible tables will be the same for two different tables and it is difficult to judge whether the both probabilities should be included in the summation or not.

Nonetheless, Fisher's exact probability test does not use any approximation and using computer software enables you to apply this test for any (even large-sized) two-way tables. Thus, usually Fisher's exact probability test is recommended to use than use of chi-square test.

In Rgui console, it is easy to apply Fisher's exact test instead of chi-square test. Simply replace `chisq.test()` by `fisher.test()`.

Exercise

For the `survey` data set, calculate the Fisher's exact probability for the actual or less probable combination of sex (`Sex`) and smoking habit (`Smoke`) when those are independent.

In Rgui console, the needed typing is as follows.

```
require(MASS)
fisher.test(xtabs(~Sex+Smoke, data=survey))
```

In Rcmdr, after activating `survey` data set, select [Statistics], [Contingency tables], [Two-way table ...], then select `Sex` as "Row variable" and `Smoke` as "Column variable", and check the box beside "Fisher's exact test" then click [OK].

Both will give the same result of `p-value = 0.3105`. The null-hypothesis cannot be rejected at 5% significance level. Thus we cannot say the siginificant association between sex and smoking habit.

9 Contingency tables for repeated measures

For ordered or categorical indices, the data for the same individuals at 2 different times can be summarized as the 2 dimensional cross table, and the table can be used to evaluate the test-retest reliability. However, neither of chi-square test nor Fisher's exact probability test is appropriate for this purpose, because the variables at 2 different times are clearly not independent. The situation is similar when the same subjects are evaluated by two different raters and it should be judged whether the inter-rater agreement is significantly more than "by chance" or not.

If you would like to know the match of 2 measurements, the Kappa statistic can be used, where the null hypothesis is that the match of 2 measurements is same as random match and the alternative hypothesis is that the 2 measurements are significantly more matching than random.

If you would like to know the effect of intervention, the null-hypothesis is same as the Kappa statistic, but the alternative hypothesis is that **the 2 measurements are more different than random match**. In such case, the McNemar test can be

applied. If the variable is ordered and number of categories is more than 3, Wilcox's signed rank test is also applicable.

9.1 Kappa statistic

Assume the next table for the evaluation of test-retest reliability.

	2nd time	
	Yes	No
1st time Yes	a	b
1st time No	c	d

If the results at 2 times completely agreed, $b = c = 0$, but usually $b \neq 0$ and/or $c \neq 0$. Here we can define the agreement probability $P_o = (a + d)/(a + b + c + d)$. When complete agreement, from $b = c = 0$, $P_o = (a + d)/(a + d) = 1$. When complete disagreement, oppositely, from $a = d = 0$, $P_o = 0$. If the extent of agreement is same as random combination, the expected agreement probability P_e can be calculated as follows: $P_e = \{(a + c)(a + b)/(a + b + c + d) + (b + d)(c + d)/(a + b + c + d)\}/(a + b + c + d)$.

If we define κ as $\kappa = (P_o - P_e)/(1 - P_e)$, $\kappa = 1$ at complete agreement, $\kappa = 0$ at the same agreement with random, and $\kappa < 0$ at more disagreement than random. Using that the variance of κ, $V(\kappa)$, is $V(\kappa) = P_e/\{(a+b+c+d)\times(1-P_e)\}$, we can calculate $\kappa/\sqrt{V(\kappa)}$, which obeys the standard normal distribution. Then we can test the null-hypothesis $\kappa = 0$ and estimate 95% confidence intervals of κ.

The additional package `vcd` provides the function `Kappa()` to calculate κ, and applying `confint()` function gives the confidence intervals. In addition, the author wrote the function to calculate κ, `Kappa.test`, which is included in the `fmsb` package. Unfortunately, Rcmdr cannot calculate κ right now.

Let's consider the numeric example. Assume the table below.

	2nd time	
	Yes	No
1st time Yes	12	4
1st time No	2	10

In Rgui console, after the installation of `fmsb` library, type as follows then you get the all results in the box below.

```
require(fmsb)
Kappa.test(matrix(c(12,2,4,10),2,2))
```

```
$Result
        Estimate Cohen's kappa statistics and test the null
        hypothesis that the extent of agreement is same as random
        (kappa=0)
data:  matrix(c(12, 2, 4, 10), 2, 2)
Z = 3.0237, p-value = 0.001248
95 percent confidence interval:
 0.2674605 0.8753967
sample estimates:
[1] 0.5714286

$Judgement
[1] "Moderate agreement"
```

Here the judgement is due to the criteria given by Landis JR, Koch GG (1977) Biometrics, 33: 159-174: If κ is less than 0, `"No agreement"`, if 0-0.2, `"Slignt agreement"`, if 0.2-0.4, `"Fair agreement"`, if 0.4-0.6, `"Moderate agreement"`, if 0.6-0.8, `"Substantial agreement"`, if 0.8-1.0, `"Almost perfect agreement"`. This is only a rough guideline, but is practically useful.

9.2 McNemar test

The original form of McNemar test is developed for 2 by 2 cross table. Assume the next table as the resulted numbers of individuals.

	after	
	Yes	No
before Yes	a	b
before No	c	d

The McNemar test calculates the χ_0^2 defined below. The χ_0^2 obeys the chi-square distribution of degree of freedom 1, under the null-hypothesis.

$$\chi_0^2 = \frac{(b-c)^2}{(b+c)}$$

If the continuity correction is applied (however, if b equals c, $\chi_0^2 = 0$), as follows.

$$\chi_0^2 = \frac{(|b-c|-1)^2}{(b+c)}$$

The extended McNemar test can be applied to M by M cross table. Let the number of individuals in cell [i,j] be n_{ij} (here i, j = 1, 2, ..., M). Calculate χ_0^2 as follows, then the χ_0^2 obeys chi-square distribution with degree of freedom being M(M-1)/2, under the null hypothesis that the probabilities of being classified into cells [i,j] and [j,i] are the same.

$$\chi_0^2 = \frac{\sum_{i<j} \left(n_{ij} - n_{ji}\right)^2}{(n_{ij} + n_{ji})}$$

In Rgui console, the McNemar test can be done by simply typing `mcnemar.test(TABLE)`, where the `TABLE` is the contingency table between corresponding 2 variables. Unfortunately, Rcmdr does not provide the function to conduct McNemar test right now.

10 Survival Analysis

The survival analyses are not supported in Rcmdr itself. However, recent versions of Rcmdr can use various plugins, among which some provide graphical interface for survival library. The name of the plugin is "RcmdrPlugin.survival" and "RcmdrPlugin.SurvivalT"[*30]. To install the former, type as follows.

```
install.packages("RcmdrPlugin.survival")
```

Fully explaining survival analysis is difficult within this introductory class, but many functions to conduct survival analysis are provided by `survival` package.

In Rcmdr, to use plugins, you need to select [Tools] in Rcmdr menu, and select [Load Plugins...], then select `RcmdrPlugin.survival` and click [OK]. The Rcmdr ask you to restart Rcmdr, then click [OK]. After restarting Rcmdr (let's refer this situation as Rcmdr+RcmdrPlugin.survival), you can find the following menus for the survival analysis: [Survival data] in [Data], [Survival analysis] and some items of [Fit Models] in [Statistics], some items of [Numerical diagnostics] and some items of [Graphs] in [Models] menu.

[*30] The former is provided and maintained by the author of Rcmdr, the latter is provided by Dr. Daniel C. Leucuta in Romania, who published an article about it as: Leucuta DC, Achimas-Cadariu A (2008) "Statistical graphical user interface plug-in for survival analysis in R statistical and graphics language and environment." *Applied Medical Informatics*, 23(3-4): 57-62.

10.1 Concept of survival analysis

In longitudinal observation of the effects by toxic substances, not only the changes of quantitative indices but also the time to event such as death could be used to evaluate the strength of the toxicity of that substances. The data like time to event can be analyzed by survival analysis (a.k.a. event history analysis).

Amongst one of the most famous methods is the Kaplan-Meier's product-limit estimate, which is the products of (1 − number of events divided by population at risk) at all times of occurences of events. The time when this value goes across 0.5 is median survival time. For the data of time to events for 2 groups, the logrank test or generalized Wilcoxon test can be used to test the difference between the 2 groups. Besides those nonparametric methods, there are parametric approaches to fit the time to events with any known distribution, like exponential distribution or Welbull distribution. The famous semi-parametric approach of survival analysis is the Cox's regression (a.k.a. fitting a proportional hazard model), which assume that the *i*th individual's hazard can be expressed as the product of the baseline hazard and $\exp(\sum \beta z_i)$, where z_i is the *i*th value of covariate vector z and β is coefficient's vector.

After typing `require(survival)` or `library(survival)`, `Surv()` generates the survival time object, `survfit()` will calculate the Kaplan-Meier's product-limit estimate (this result can also be used to draw survival curves), `survdiff()` will conduct the logrank test, and `coxph()` fit the proportional hazard model to the data. If you want to know the survival analysis in detail, you should read another text like Bull *et al.*, 1997.

10.2 Kaplan-Meier method

Let the times of event happening since beginning of time at risk $t_1, t_2, ...$, the numbers of events at each time $d_1, d_2, ...$, and the size of population at risk just before the each time $n_1, n_2, ...$ The size of population at risk decreases not only by the event occurrence but also by censoring such as moving out or loss to follow up or death by competing risks. When the censoring and event occurred at the same time, usually the censoring occurred just after the event happening.

Here the Kaplan-Meier's product-limit estimates $\hat{S}(t)$ can be defined as follows.

$$\hat{S}(t) = (1 - d_1/n_1)(1 - d_2/n_2)... = \prod_{i<t}(1 - d_i/n_i)$$

Clearly this value means the probability of survival and is numerically 1 at first (nobody has experienced the event) and 0 in the end (after the everybody experienced the event).

The standard error of $\hat{S}(t)$ is given by the Greenwood's formula shown below.

$$\mathrm{var}(\hat{S}) = \hat{S}^2 \times \sum_{i<t} \frac{d_i}{n_i(n_i - d_i)}$$

Estimated $\hat{S}(t)$ is usually plotted as survival curve with 95% confidence intervals[*31].

In Rgui console, basic grammer for Kaplan-Meier method is shown below (with comments). After loading survival package in memory by typing `library(survival)` or `require(survival)`, `dat <- Surv(times, flags)` generates the survival time data `dat`, where the flags for censoring become 1 when the observation ends by event's occurrence and become 0 when the observation ends by censoring. Conducting Kaplan-Meier method is simply `res <- survfit(dat~1)`, or `res <- survfit(dat~group)` if you want to estimate this by group. By typing `plot(res)`, we can get the graph of survival curves. Detailed result of estimation can be obtained by `summary(res)`.

[*31] However, if $\hat{S}(t)$s were estimated for 2 groups and drawing both to compare them, confidence intervals are not usually drawn.

practice

The aml data.frame in the survival package is the result of randomized controlled trial for the maintenance chemotherapy's effect to delay the remission of acute myelogenous leukemia. Included variables are the following 3.

time time to remission or censoring in weeks.

status flag for censoring, where 0 means censoring, 1 means remission.

x whether maintenance chemotherapy was conducted or not (Maintained or Nonmaintained).

Let's conduct Kaplan-Meier's method to estimate survival curves for the 2 groups (under maintenance chemotherapy or not). How long the median survival times (here, median times to remission) of the 2 groups are?

In Rgui console, type as follows.

```
require(survival)
print(res <- survfit(Surv(time, status)~x, data=aml))
plot(res, xlab="(Weeks)", lty=1:2, main="Periods until remission of acute myelogenous leukemia")
legend("right", lty=1:2, legend=levels(aml$x))
```

The second line conducts Kaplan-Meier estimation and shows result below.

```
                records n.max n.start events median 0.95LCL 0.95UCL
x=Maintained         11    11      11      7     31      18      NA
x=Nonmaintained      12    12      12     11     23       8      NA
```

The number of maintained patients is 11 and non-maintained patients is 12. Among them, remission cases were 7 and 11, respectively. Median periods until remission were 31 weeks for maintained patients and 23 weeks for non-maintained patients. Lower limits of 95% confidence intervals of the median periods until remission were 18 weeks and 8 weeks, respectively. Upper limits of those were both infinity. The third line draws survival curves as solid line and dashed line for maintained group and non-maintained group, respectively. The fourth line adds a legend to the graph.

In Rcmdr+RcmdrPlugin.survival, first of all, activate the dataset aml **in** survival **package. Select [Data], [Data in packages], [Read data set from an attached packages...], then double-click** survival **in the top-left box, subsequently double-click** leukemia **in the top-right box (Actually** aml **and** leukemia **are the same, but the former will not appear in this box). After that, click [OK].**

Kaplan-Meier estimate will be done by select [Statistics], [Survival analysis], [Estimate survival function], then select [time] as "Time or start/end times", select [status] as "Event indicator". As the "Strata", [x] should be selected if you would like to do Kaplan-Meier estimate separately for maintained group and non-maintained group, otherwise leave there unselected.

This menu can draw the survival curve. The options for "Confidence Intervals" can be selected from "Log", "Log-log", "plain", and "none". The default is "Log", because the default value of conf.type= **option in** survfit() **function is** conf.type="log"**. In the Japanese textbook by Ohashi and Hamada (1995), SAS version 6's default is same as** conf.type="plain"**. The tutorial paper by Bull and Spiegelhalter (1997) showed the same result with** conf.type="log-log"**. The options for "Plot confidence intervals" can be selected from the [Yes] (always drawing confidence intervals of survival curve), [No] (always not drawing them), and [Default] (drawing them only when "Strata" is unspecified). The default value of "Confidence level" is .95 (95%). Typing [.99] there leads to draw survival curve with 99% confidence intervals. "Mark censoring times" is checked by default, which means to plot tickmarks at the time of censoring.**

There are some other options in "Method" and "Variance Method" than [Kaplan-Meier] and [Greenwood], but I recommend to leave there as default. The box besides "Quantiles to estimate" is [.25,.5,.75] by default, which should also be left unchanged. The [.5]'s point of output window shows median time to event occurrence. After specifying all of them, click [OK], then you see the results.

10.3 Logrank test

Here I will give a brief explanation about the concept of logrank test. Let's imagine 8 rats and randomly assign 2 groups (administering toxic substance A and B) to them, then follow them up. On the day 4, 6, 8, 9, the rats in the first group (which took toxic substance A) died. On the day 5, 7, 12, 14, the rats in the second group (which took toxic substance B) died. There is no censoring.

The concept of logrank test is, making 2 by 2 contingency tables of group and alive/dead at each time of event, and calculate the common chi-square value in a Cochran-Mantel-Haentzel's manner.

In the example shown above, let's denote expected number of death for jth group at the ith time of events as e_{ij}, observed total number of death at time i as d_i, population at risk of jth group at time i as n_{ij}, total population at risk at time i as n_i, then

$$e_{ij} = d_i \cdots n_{ij}/n_i$$

In the above example, $e_{11} = 1 \cdots 4/8 = 0.5$. Next, denote the number of death of jth group at time i as d_{ij}, the weight of time i as w_i, the score of jth group at time i as u_{ij}, then

$$u_{ij} = w_i \cdots (d_{ij} - e_{ij})$$

In the logrank test, every weights are 1. Then

$$u_{ij} = d_{ij} - e_{ij}$$

The summary score for group j, u_j can be calculated as follows.

$$u_j = \sum_i d_{ij} - e_{ij}$$

The variance V of the score can be considered as follows.

$$V = V_{jj} = \sum_i \frac{(n_i - n_{ij})n_{ij} \cdots d_i(n_i - d_i)}{{n_i}^2(n_i - 1)}$$

The chi-square value ${\chi_0}^2$, which obeys chi-square distribution of degree of freesom 1, is given by the formula below.

$${\chi_0}^2 = {u_1}^2/V$$

In the above example, $u_1 = (1-4/8)+(0-3/7)+(1-3/6)+(0-2/5)+(1-2/4)+(1-1/3)+(0-0/2)+(0-0/1) = 1.338...$ and $V = 1.568...$, then ${\chi_0}^2 = 1.338^2/1.568 = 1.14$. Because 1.14 is much smaller than 3.84, which is 95% point of chi-square distribution with degree of freedom 1, we cannot judge that there is significant difference between the two groups.

In Rgui console, the script conducting this is very simple as follows.

```
require(survival)
time <- c(4,6,8,9,5,7,12,14)
event <- rep(1,8)
group <- c(1,1,1,1,2,2,2,2)
survdiff(Surv(time,event) ~ group)
```

In the case of `aml` dataset in `survival` package, the script will be as follows.

```
require(survival)
survdiff(Surv(time, status) ~ x, data=aml)
```

The resulting p value is 0.0653, which means not significant difference between the maintained group and non-maintained group at the significance level of 5%.

In Rcmdr+RcmdrPlugin.survival, logrank test of the null-hypothesis that remission time are not different between maintained and nonmaintained groups can be conducted as follows. First, activate `leukemia` (as already mentioned above). Second, select [Statistics], [Survival analysis], [Compare Survival Functions]. A window appears, then click [time] at the box of "Time or start/end times", click [status] at the box of "Event indicator", click [x] at the box of "Strata". Leave "rho" 0 as is, then click [OK]. You can get the result of logrank test in the output window.
Note: if you set 1 at the box of "rho", the generalized Wilcoxon test in a manner of Peto-Peto instead of logrank test is conducted.

10.4 Cox regression

The Kaplan-Meier estimate and logrank test assume no specific distribution in the population, thus nonparametric method. Cox regression assumes that the individuals' hazards are "proportional" to the common baseline hazard. In that sense, it's semi-parametric method.

The basic idea of the Cox regression is: Denote the covariates' vectors affecting the occurrence of events as $z_i = (z_{i1}, z_{i2}, ..., z_{ip})$ for ith individual. Denote the instantaneous rate of the event occurrence for this individual at time t as $h(z_i, t)$. This is called as "hazard function". Cox regression assumes the following formula.

$$h(z_i, t) = h_0(t) \cdot \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + ... + \beta_p z_{ip})$$

where $h_0(t)$ is the common baseline hazard, which is the instantaneous rate of event occurrence at time t of "base individual", who has no effect on event occurrence by all covariates. Unknown parameters $\beta_1, \beta_2, ..., \beta_p$ should be estimated. The effect of covariates on event occurrence is the proportional coefficients as $\exp(\beta_x z_{ix})$. This is called as "proportional hazard".

The original model by Cox considered the time-dependent covariates where z_i changes by time. However, usually we assume the effect of covariates on the event occurrence is independent from time (thus not changing by time). Therefore, the ratio of hazards between different individuals is constant regardless with time: The ratio of the 1st individual's hazard at time t to the 2nd individual's hazard at time t is not including $h_0(t)$ (cancelled from numerator and denominator), then the hazard ratio is given by the following formula. It means that the hazard ratio doesn't depend on the shape of $h_0(t)$.

$$\frac{\exp(\beta_1 z_{11} + \beta_2 z_{12} + ... + \beta_p z_{1p})}{\exp(\beta_1 z_{21} + \beta_2 z_{22} + ... + \beta_p z_{2p})}$$

The relationship of survival function and hazard function can be summarized as follows. Denote the non-negative random variable showing the time to event occurrence as T, then the survival function $S(t)$ is the probability of $T \geq t$. By this definition, $S(0) = 1$. The hazard function $h(t)$ is the instantaneous probability of event occurrence at time t. Then we can get the following equations.

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} \\ &= -\frac{dS(t)}{dt} \frac{1}{S(t)} = -\frac{d(\log(S(t))}{dt} \end{aligned}$$

The cumulative hazard function $H(t)$ is,

$$H(t) = \int_0^t h(u)du = -\log S(t)$$

Thus we obtain

$$S(t) = \exp(-H(t))$$

Denote the cumulative hazard function at time t of an individual with covariates vector z as $H(z,t)$, the survival function of the same individual as $S(z,t)$. If the proportional hazard stands,

$$H(z,t) = \int_0^t h(z,u)du = \int_0^t h_0(u)\exp(\beta z)du = \exp(\beta z)H_0(t)$$

Then we obtain the following equation.

$$S(z,t) = \exp(-H(z,t)) = \exp\{-\exp(\beta z)H_0(t)\}$$

Take the logarithm and inverse the sign and take the logarithm again, the we obtain the following equation.

$$\log(-\log S(z,t)) = \beta z + \log H_0(t)$$

From this equation, we can see that the pararell curves with the gap of βz will be drawn, being survival time as horizontal axis and $\log(-\log S(z,t))$ as vertical axis. If this pararell nature is not met, the "proportional hazard" assumption cannot stand and thus Cox regression is not suitable.

In estimation of β, the concept of partial likelihood is applied. If we decompose the probability of event occurring for ith individual at time t into the probability of single event occurrence at time t and the conditional probability that the event occurred for the specific individual i under the condition that the event occurred at time t, the former is still unknown unless we can assume any specific parametric model, but the latter $L(i,t)$ can be always estimated as the ratio of ith individual's hazard being numerator and the sum of all individuals' hazard within the whole population at risk at time t.

For the all event occurrences, denote the products of all $L(i,t)$s as L, the meaning of L is the total likelihood minus the likelihood concerning time, thus is called as partial likelihood. To estimate a "good" parameter β that asymptotically converges to the true parameter as the sample size becomes larger, and whose distribution obeys a normal distribution, and whose variance becomes smallest, Cox conjectured that such β could be obtained when the L became maximum and this conjecture was given prove by the Martingale theory. By this fact, the proportional hazard model is also known as Cox regression[*32]. The basic form of Cox regression in R is `coxph(Surv(time,cens)~grp+covar,data=dat)`.

Practice

In the `aml` dataset, conduct Cox regression on the effect of treatment (maintained / nonmaintained) on the survival time.

```
require(survival)
summary(res <- coxph(Surv(time,status)~x, data=aml))
KM <- survfit(Surv(time,status)~x, data=aml)
par(family="sans",las=1,mfrow=c(1,3))
plot(KM, lty=1:2, main="Kaplan-Meier plot of survival time of aml dataset.")
legend("topright", lty=1:2, legend=levels(aml$x))
plot(survfit(res),
 main="The survival curve of the reference individual\n with the treatments being covariates")
plot(KM, fun=function(y) {log(-log(y))}, lty=1:2, main="Double logarithmic plot of aml dataset")
```

The result given in the second line is shown below.

[*32] If multiple events occur simultaneously, there are several methods to treat them: Exact method, Efron's method, Breslow's method, discrete method, and so on. However, whenever possible, Exact method is recommended. The discrete method should be used when the survival times are given as discrete measures. Many statistical software uses Breslow's method, but the default method in R's `coxph()` function is Efron's method. Generally speaking, Efron's method gives closer results than Breslow's method.

```
Call:
coxph(formula = Surv(time, status) ~ x, data = aml)

  n= 23
                coef exp(coef) se(coef)    z     p
xNonmaintained 0.916       2.5    0.512 1.79 0.074

               exp(coef) exp(-coef) lower .95 upper .95
xNonmaintained       2.5        0.4     0.916      6.81

Rsquare= 0.137   (max possible= 0.976 )
Likelihood ratio test= 3.38  on 1 df,   p=0.0658
Wald test            = 3.2  on 1 df,   p=0.0737
Score (logrank) test = 3.42  on 1 df,   p=0.0645
```

The p value of the test of null-hypothesis that the maintained and nonmaintained group have the same hazard is 0.074, so that the null-hypothesis is not rejected at the 5% significance level[*33]. The `exp(coef)` 2.5 is the estimated hazard ratio of 2 groups, so that we can judge the nonmaintained group has 2.5 times higher hazard of maintained group's hazard but the 95% confidence intervals includes 1.

By the third line and later scripts, 3 graphs are drawn. From left to right, the Kaplan-Meier plot estimated for 2 groups separately, the baseline survival curve with 95% confidence intervals as the result of Cox regression with treatment being covariates, and the double logarithmic plot are drawn, respectively.

If you dare to draw the baseline survival curves of Cox regression with covariates for the 2 treatment groups separately, for example, `subset=(x=="Maintained")` option can be used in the `coxph()` function, when the group variable cannot be included as covariates. More than 2 survival curves could be drawn without erasing previous graphs by specifying `par(new=TRUE)` option. However, I don't recommend this manner.

There are three strategies to control the effect of covariates on the survival time. For example, to analyze the survival time of cancer patients, the effects of stage should be controlled. The possible strategies to control them are:

1. Analyzing the survival time separately by each stage.
2. Assuming that the effects of other covariates are common to all stages, then set the stage as strata.
3. Including the stage as covariates in the same model.

The third strategy has an advantage that the effect of stage can be quantitatively estimated, but it requires the unrealistic condition that the baseline hazards are the same for all stages. In addition, the method of coding stages as covariates may affect the result (usually coding as dummy variable).

The second strategy means that the baseline hazards are different by stage. In the `coxph()` function, the option `strata()` can be used for different baseline hazards. For example, if the data.frame of survival time of cancer patients is `aml`, which includes 4 variables: the variable of survival time `time`, the variable of censoring flag `status`, the group variable showing the treatment `x`, the variable of stage of cancer progress `stage`, then the model can be written as `coxph(Surv(time,status)~x+strata(stage), data=aml)`[*34]

Anyway, Cox regression is a kind of model-fitting, so that we can select better models by the residual analysis, likelihood ratio test, and the squared multiple correlation coefficients, but AICs are usually not available because calculating AIC requires the specified distribution in the baseline hazard.

[*33] `Score (logrank) test` in the bottom line is the result of Rao's score test, different from the logrank test with `survdiff()`.

[*34] However, in fact, `aml` data.frame does not include stage.

In Rcmdr+RcmdrPlugin.survival, select [Statistics], [Fit models], [Cox regression model], then select `time` in the box of "Time or start/end times", select `status` in the box of "Event indicator", and remain "Strata" and "Clusters" unselected (if you select "Strata", different baseline hazards by strata are assumed). Next, select `Exact` in the box of "Method for Ties" (though default is `Efron`). As the box of "Robust Standard Errors", any of `Default`, `Yes`, or `No` is selectable (usually should be remained as `Default`). After specifying all those above, write the group variables connecting with covariates by `+` in the box of "Variables". In the aml dataset, double-click `x [factor]`. After all, click [OK], then you get the result.

11 Report theme

The data `http://phi.med.gunma-u.ac.jp/grad/worldfactbook2011.txt` is a tab-delimitted text file, which is originally published as "The world factbook 2011" (CIA) at the following url.

`https://www.cia.gov/library/publications/the-world-factbook/index.html`

The variables included there are:

COUNTRY	The name of the country
IMR	Infant mortality rates (the number of death in age 0 per 1,000 live births)
LIFEEXP	Life expectancy at birth in years
TFR	Total Fertility Rates
NDAIDS	Number of death within a year caused by HIV/AIDS
APHIVAIDS	Adult prevalences in percent living with HIV/AIDS
GDPPCUSD	Gross Domestic Products per capita based on purchasing power parity in US dollar
PUNEMP	Proportion of unemployed individuals within a labor aged population in percent
GINI	Gini index measuring the degree of inequality in the distribution of family income in a country (perfect equality showing 0, perfect inequality showing 100)

Recently, many researchers of social epidemiology indicated that health levels of the country are affected by its social settings; for instance, life expectancy at birth is affected by income inequality and by gross domestic products per capita. Let's analyze this data from a such viewpoint of social epidemiology.

At first, you must **make graphs for each variable and calculate descripitve statistics** and **formulate working hypothesis** about the relationship between health levels and socioeconomic factors. After that, **select and conduct the appropriate methods** to analyze the working hypothesis. **Properly describe the results (making table is recommended)** of the analyses, and **fully discuss your hypothesis** based on the results with references (if any).

Please submit the report to Minato Nakazawa via e-mail (nminato@med.gunma-u.ac.jp) until 30 June 2011, when the subject of the mail must be `[REPORT MEDICAL STAT GUNMA UNIV]`. The report should be attached as a document file within 3 pages in A4, made by Microsoft Word or OpenOffice.org Writer, or as a PDF file.

12 Furthur Readings

- Armitage P, Berry G, Matthews JNS (2002) *Statistical Methods in Medical Research, 4th ed.*, Blackwell Publishing.
- Bull K, Spiegelhalter DJ (1997) Tutorial in biostatistics: Survival analysis in observational studies. *Statistics in Medicine*, 16: 1041-1074.
- Faraway JJ (2006) *Extending the linear models with R: Generalized linear, mixed effects and nonparametric regression models.* Chapman and Hall.
- `http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf` (The introductory textbook for R Commander, provided by Prof. John Fox (McMaster Univ.), the developper of Rcmdr package.)
- Maindonald J, Braun J (2003) *Data analysis and graphics using R*, Cambridge Univ. Press.
- Nagelkerke N (1991) A note on a general definition of the coefficient of determination. *Biometrika*, 78: 691-692.
- R Development Core Team (2010). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.
- Venables WN, Ripley BD (1999) *Modern Applied Statistics with S-PLUS. Third Edition.* Springer.