データ解析

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

この講義について

講義の目的

- フリーの統計解析用言語「R」を用いてデータ解析の仕方を学ぶ.
- 実際に自分で手を動かすことによってデータ解析手法を習得.
- そのため頻繁にレポート提出をしてもらう.

成績評価

- min(出席 30% + レポート 80%, 100)

前提知識

- 確率統計の基本的な知識があることが望ましい.
- 前学期の確率統計第二を取っていれば問題なし.

R とは

- オープンソース・フリーソフトウェア の統計解析向けのプログラミング言語及びその開発実行環境.
- R の使い方に慣れておけば他の言語にも活用可能.
  - 類似言語：Matlab, Octabe, Python
- 多くの推定・検定方法が実装されていて，誰でも簡単に統計解析ができる.

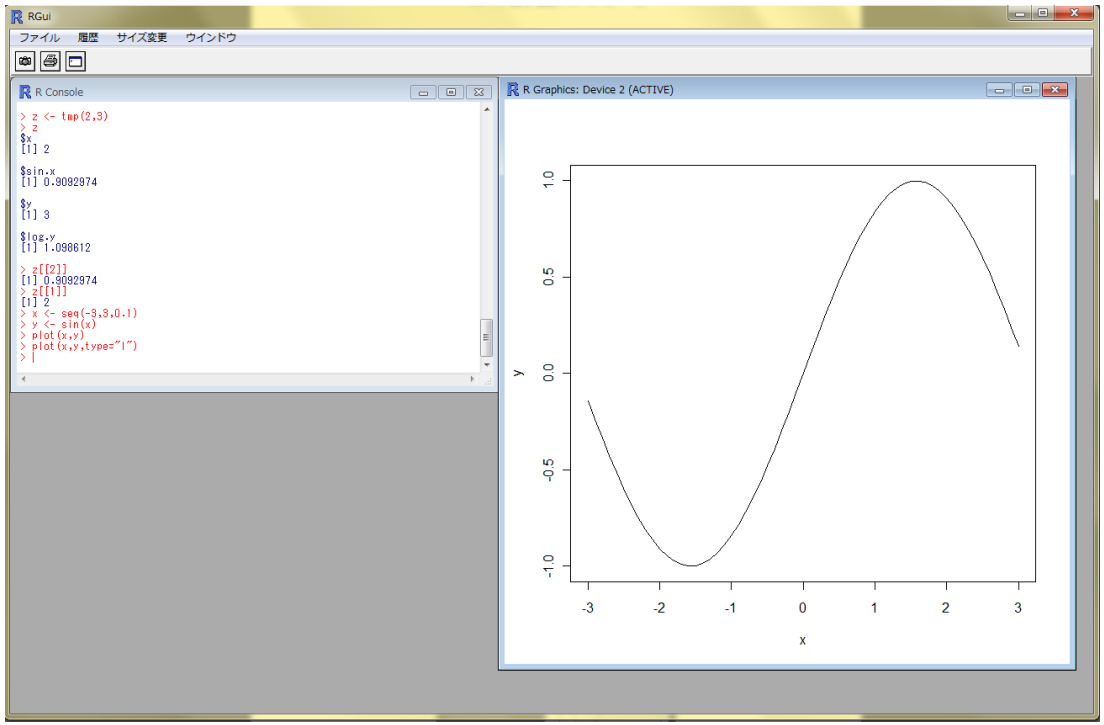


Figure: R の操作画面

講義内容（予定）

- Ｒの基本的操作
- 確率変数，確率分布の生成
- 回帰分析
  - 単回帰，重回帰
  - ガウスマルコフモデルにおける検定
- 判別分析
  - Fisher 線形判別分析
  - ロジスティック回帰，サポートベクトルマシン
- 検定
  - 適合度検定
  - 独立性検定
  - 2 標本検定
- 主成分分析
- ノンパラメトリック推定
  - カーネル密度推定
  - カーネル平滑化回帰
- クラスタリング・トピックモデル
- 時系列解析

回帰分析

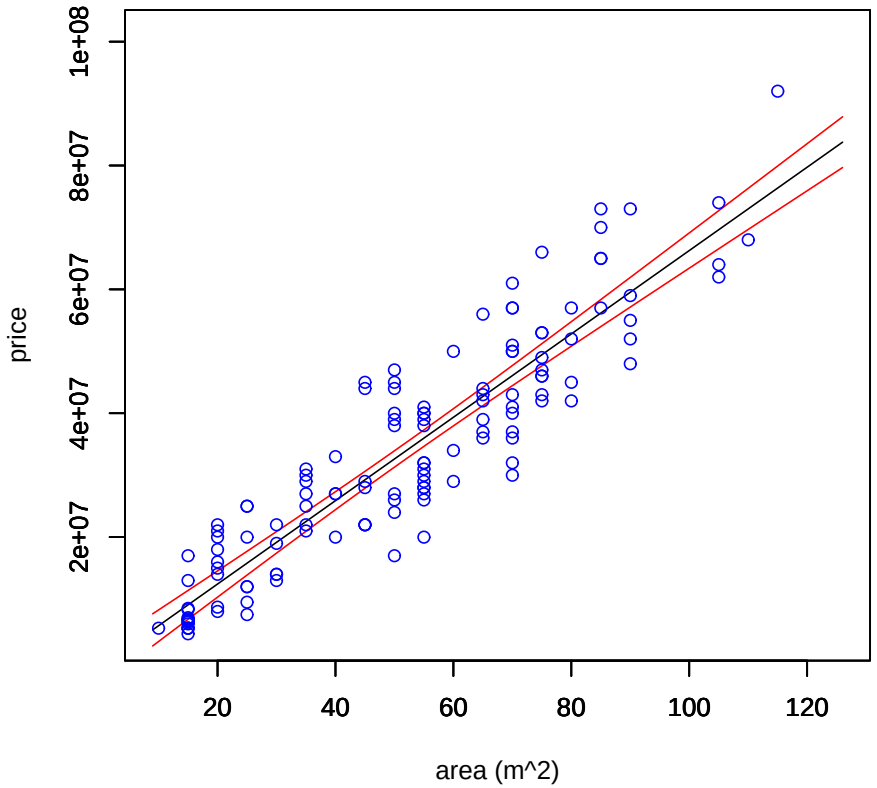


Figure: 床面積とマンション価格

判別分析

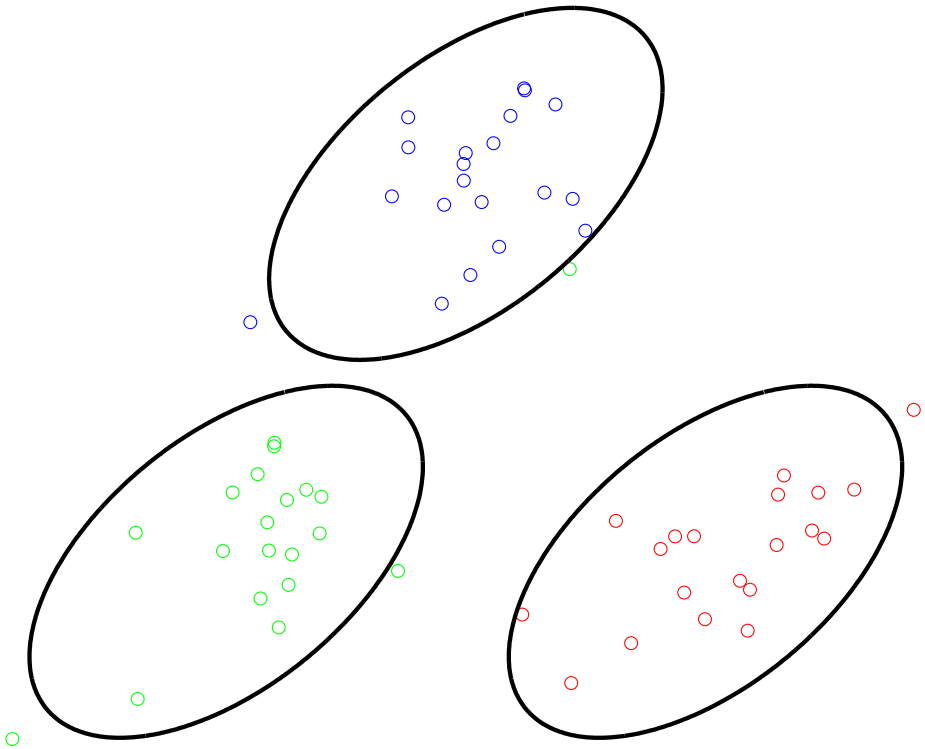


Figure: 線形判別分析

検定

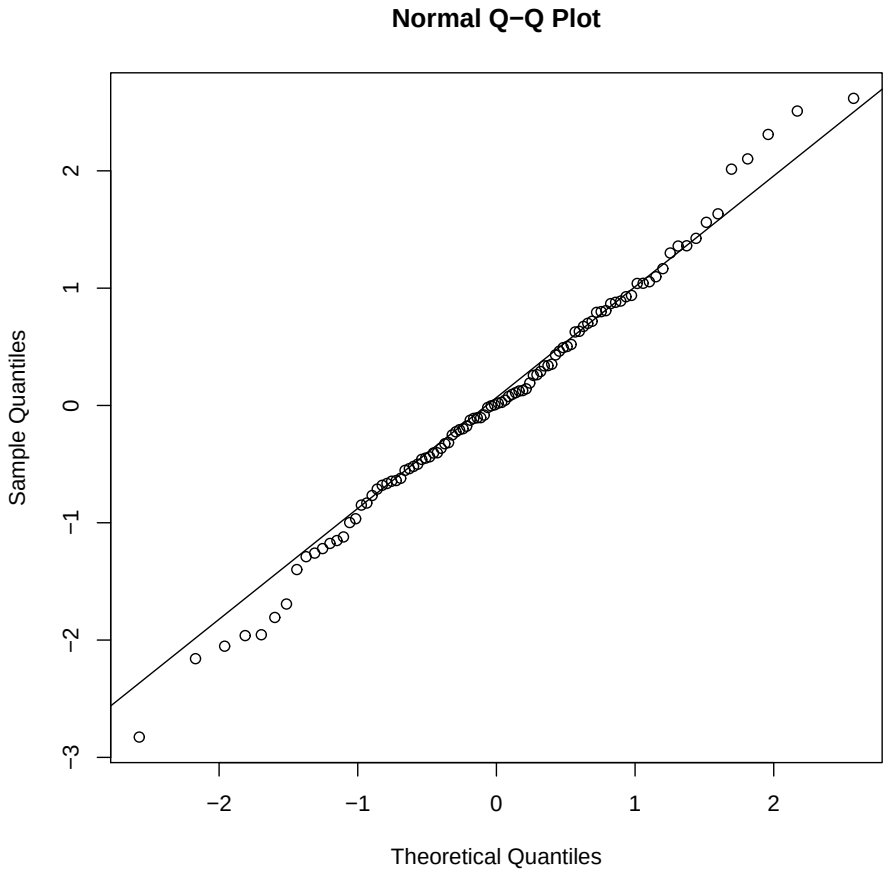


Figure: 正規性検定

ノンパラメトリック密度推定

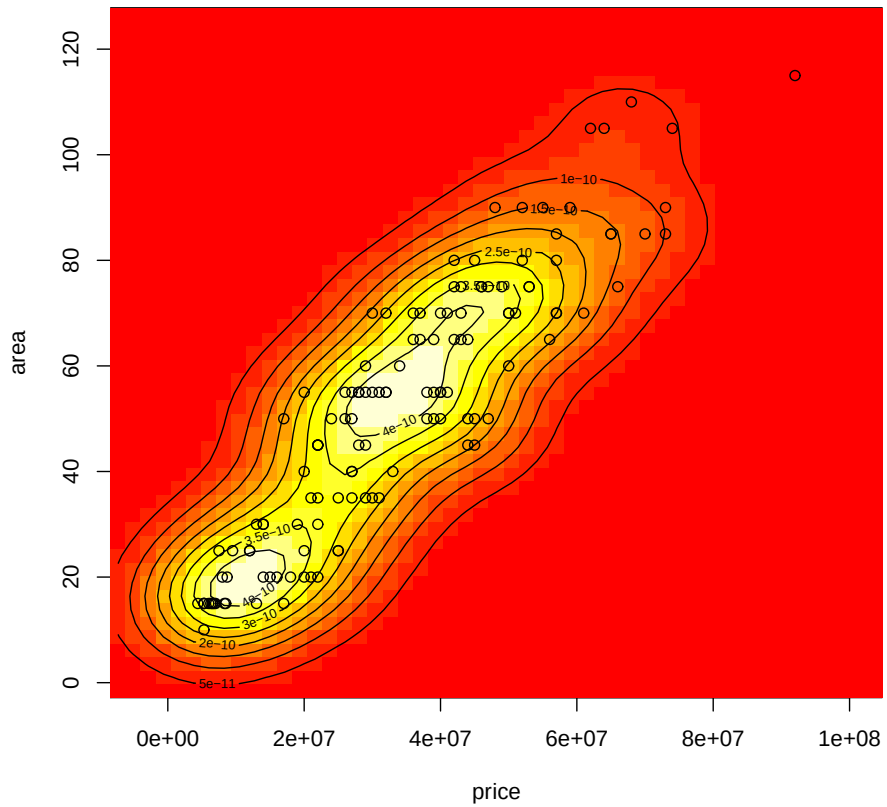
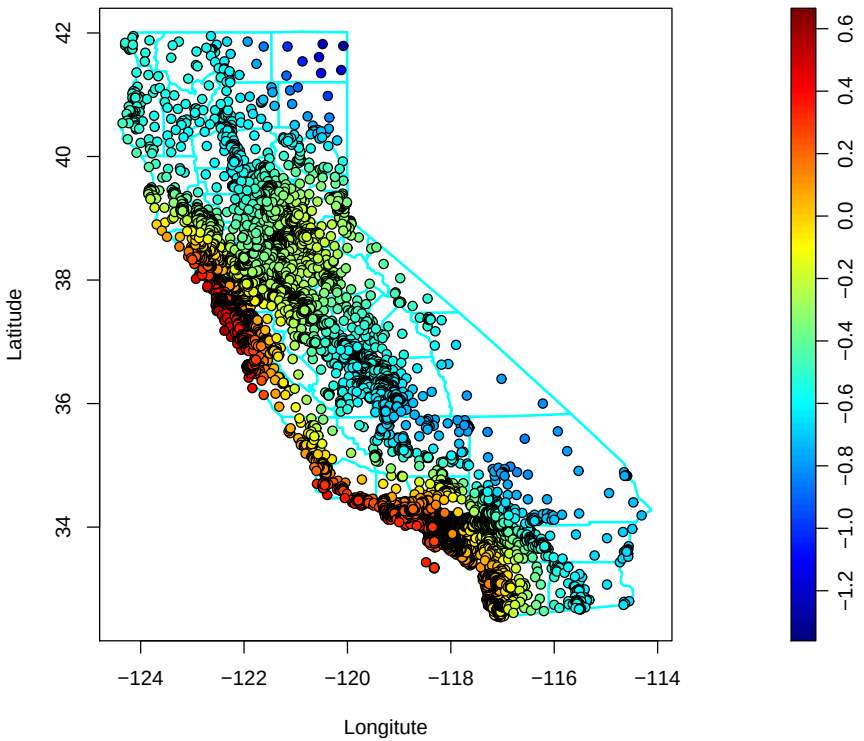
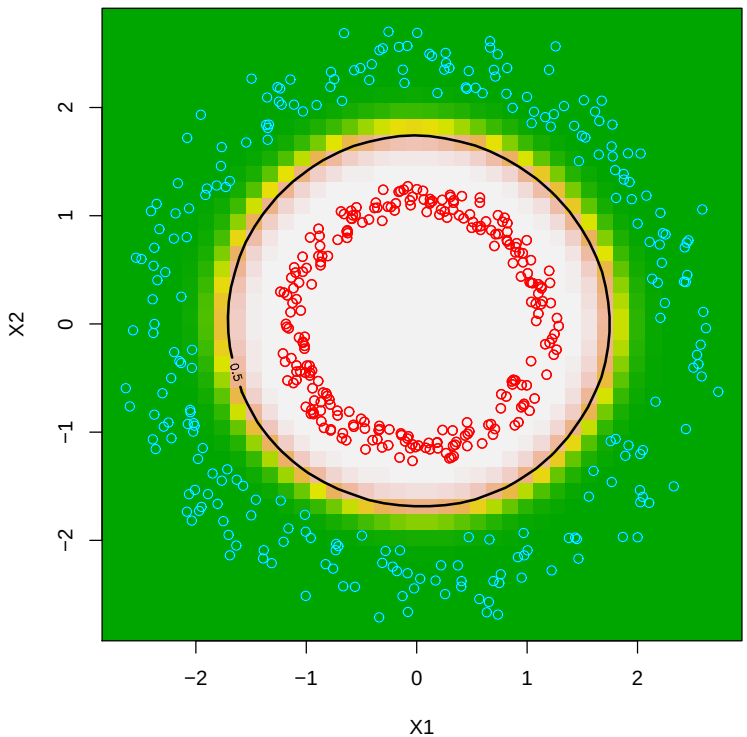


Figure: マンション価格と床面積の分布

ノンパラメトリック推定



(a) 加法モデル（カリフォルニア州の住宅価格）



(b) ノンパラメトリック判別分析

クラスタリング・トピックモデル

「Wikipedia の記事分類」

"Topic 3 :"
"架空の国一覧 | 岡村明美 | 佐久間レイ | 三木眞一郎 | 石田彰 | うえだゆうじ | 山口勝平 | 根谷美智子 | 広瀬正志 | 小西克幸 | 八奈見乗児 | 山口由里子 | 進藤尚美 | くまいもとこ | 関俊彦 | 千葉一伸 | 草尾毅 | 坂本千夏 | 飛田展男 | 三宅健太"

"Topic 4 :"
"Xeon | PC-9821シリーズ | 順序数 | ThinkCentre | Safari | Microsoftオン化傾向 | X68000 | Unicode一覧 E0000-E0FFF | MC68000 | .NET Framew

"Topic 18 :"
"中国帝王一覧 | 元号一覧（日本） | 天文（元号） | 従一位 | 後白河天皇 | 延暦 | 享保 | 紀元前1千年紀 | 文化（元号） | 伺候席 | 征夷大将軍 | 夏商周年表 | 宝暦 | 備前国 | 守護代 | 醍醐天皇 | 伊勢国 | 摂津国 | 相模国 | 紀元前4世紀"

今日の講義内容

- R ことはじめ
- ベクトル操作
- 行列操作
- データフレーム
- リスト

R のインストール

R を Windows にインストールする方法.

`http://cran.r-project.org/`

へ行き,
「Windows」→「base」→「Download R (version No.) for Windows」
とクリックしゆく. すると `R-(version No.)-win32.exe` のダウンロードが始まるのでこれを「実行」. 後はデフォルトの設定を利用すれば問題なくインストールできる. 一番最初に言語の選択を尋ねられるので「Japanese」を選べば日本語環境を利用できる.

R for Windows

Subdirectories:

base — Binaries for base distribution (managed by Duncan Murdoch). This is what you want to **install R for the first time**.

contrib — Binaries of contributed packages (managed by Uwe Ligges). There is also information on third party software available for CRA Windows services and corresponding environment and make variables.

Rtools — Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions binaries.

You may also want to read the R FAQ and R for Windows FAQ.

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

↓

R-3.0.3 for Windows (32/64 bit)

Download R 3.0.3 for Windows (54 megabytes, 32/64 bit)

Installation and other instructions

New features in this version

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the md5sum of the .exe fingerprint. You will need a version of md5sum for windows: both graphical and command line versions are available.

Frequently asked questions

- How do I install R when using Windows Vista?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

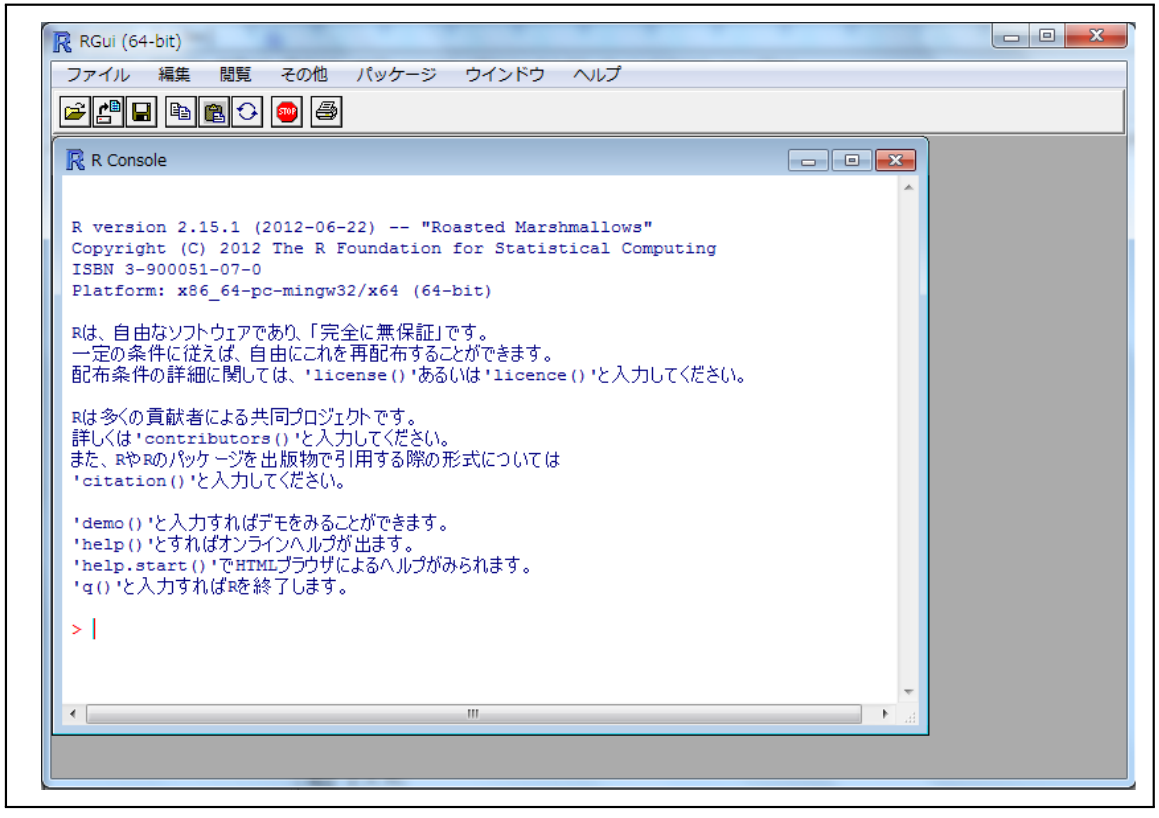
Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Other builds

- Pre-release test versions are available here.
- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.

R の起動と終了

- R の起動画面：



- R の終了：`q()` と入力.

基本的なコマンド

- 関数のヘルプ：help(’関数名’)
  例：help(’rnorm’)
- ワーキングディレクトリの表示： getwd()
- ワーキングディレクトリの移動： setwd("ディレクトリパス")
  ※ バックスラッシュをパス区切りに使う場合は，\\のように二回ずつ入れる．
  × setwd("C:\Users\.."), ○ setwd("C:\\Users\\...")
- パッケージのインストール：install.packages("パッケージ名")
  例：install.packages("Rcmdr")
  ※ ミラーサイトの指定を求められたら「Japan (Tokyo)」など近場を選べばよい．
- パッケージ（ライブラリ）の読み込み： library(ライブラリ名)
  例：library(Rcmdr)
- スクリプトファイルの実行： source("hoge.R")
  ※ hoge.R の中に書いてあるコマンドが実行される．
  ※ 関数の定義を書いておけば関数を読み込むことができる．

Rのデータタイプ

- 数値

```
x <-1.5      # 数値（倍精度実数）
```

- ベクトル

```
x <- c(1,2,3)  #数値ベクトル
```

- 行列

```
(x <- matrix(1:12,nrow=4,ncol=3)) # 行数と列数を指定
```

- データフレーム

```
(x <- data.frame(a=1:4,b=runif(4),c=month.abb[1:4])) # 数値・数値・文字
```

- リスト

```
(x <- list(a=1:3,b=rnorm(5),c=letters[1:6],d=c(sin,cos,log,exp))
```

ベクトルの操作

```
> x <- c(1,2,3)
> x + 0.1
[1] 1.1 2.1 3.1
> 2*x
[1] 2 4 6
> x+x    #要素毎の和
[1] 2 4 6
> x*x    #要素毎の積
[1] 1 4 9
> x^3    #要素ごとのべき
[1]  1  8 27
> y <- c(5,3,4)
> x + y
[1] 6 5 7
```

行列の操作

```
> x <- matrix(1:12,nrow=4,ncol=3)
> x
     [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
> t(x)
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
> x[2,3]              # (2,3) 要素
[1] 10
> x[,3]               # 第 3 列
[1]  9 10 11 12
> x[2,]               # 第 2 行
[1]  2  6 10
```

行列の演算

```
> (x <- matrix(1:8,ncol=4,nrow=2)
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
> (y <- matrix(runif(8),ncol=4,nrow=2))
          [,1]       [,2]      [,3]      [,4]
[1,] 0.3997917 0.69898767 0.4798853 0.7412236
[2,] 0.5379282 0.06090523 0.2389898 0.9024541
> x %*% t(y)          # 行列積
         [,1]     [,2]
[1,] 10.08475 8.232771
[2,] 12.40463 9.973049
```

行列の演算

```
> x <- matrix(rnorm(9),nc=3,nr=3)
> det(x)                                          # 行列式
[1] -1.757556
> solve(x,diag(3))                                # 逆行列
           [,1]       [,2]       [,3]
[1,]  0.7420844  0.2275223  0.1775263
[2,] -0.7177414 -1.3963191 -0.1661068
[3,]  0.9885972  2.0252123  0.8801354
> solve(x)                                        # これでもＯＫ
           [,1]       [,2]       [,3]
[1,]  0.7420844  0.2275223  0.1775263
[2,] -0.7177414 -1.3963191 -0.1661068
[3,]  0.9885972  2.0252123  0.8801354
> x%*%solve(x)                                    # ほぼ単位行列
              [,1]          [,2]          [,3]
[1,] 1.000000e+00  0.000000e+00 0.000000e+00
[2,] 1.257675e-16  1.000000e+00 1.040834e-17
[3,] 0.000000e+00 -4.440892e-16 1.000000e+00
```

行列の分解

```
> x <- matrix(runif(3*3),ncol=3,nrow=3)
> (res <- eigen(x)) # 固有値分解
$values
[1] 1.9863367+0.0000000i 0.2288255+0.1546637i 0.2288255-0.1546637i

$vectors
             [,1]                    [,2]                    [,3]
[1,] 0.6782596+0i  0.7406643+0.0000000i  0.7406643+0.0000000i
[2,] 0.5053572+0i -0.5153904+0.2999935i -0.5153904-0.2999935i
[3,] 0.5334585+0i  0.1440272-0.2739512i  0.1440272+0.2739512i

> (x%*%res$vectors) - (res$vectors%*%diag(res$values))
  # 確かに固有値固有ベクトル分解
                  [,1]                          [,2]
[1,] -1.332268e-15+0i -1.942890e-16+2.775558e-17i
[2,] -8.881784e-16+0i  8.326673e-17-6.938894e-17i
[3,] -1.110223e-15+0i -2.775558e-17-6.938894e-18i
                              [,3]
[1,] -1.942890e-16-2.775558e-17i
```

特異値分解

```
> (res <- svd(x)) # 特異値分解
$d
[1] 2.122693 0.496124 0.143879

$u
           [,1]       [,2]       [,3]
[1,] -0.6286352  0.7498205 -0.2063660
[2,] -0.5604081 -0.6207376 -0.5482952
[3,] -0.5392221 -0.2290285  0.8104230

$v
           [,1]       [,2]       [,3]
[1,] -0.3700176  0.7911143  0.4870577
[2,] -0.6632158  0.1421883 -0.7347974
[3,] -0.6505626 -0.5949123  0.4720673

> norm(x - (res$u %*% diag(res$d) %*% t(res$v)))  # 確かに特異値分解
[1] 7.771561e-16
```

データフレーム

- データフレームとは異なる型のベクトルをまとめて一つの変数として扱える配列.
- 見た目は行列と同じ．異なる型を各列に入れられる点が違う.

```
> (x <- data.frame(a=1:4,b=runif(4),c=month.abb[1:4])) # 数値・数値・
  a         b   c
1 1 0.1068124 Jan
2 2 0.5287629 Feb
3 3 0.3822495 Mar
4 4 0.4525498 Apr
> x[[1]]      # 第一変数を取り出し
[1] 1 2 3 4
> x$a         # 第一変数を変数名で取り出し
[1] 1 2 3 4
> x$a[2:3]
[1] 2 3
> x[1,3]
[1] Jan
Levels: Apr Feb Jan Mar
```

```
> x[,3]
[1] Jan Feb Mar Apr
Levels: Apr Feb Jan Mar
> x[3,]                          # 第 3 ケースを取り出す
  a         b   c
3 3 0.3822495 Mar
```

リスト

- ベクトルを集めたもの．各列の要素数がバラバラでも良い．

```
> (x <- list(a=1:3,b=rnorm(5),c=letters[1:6],d=c(sin,cos,log,exp)))
$a
[1] 1 2 3

$b
[1]  1.286706 -1.009312 -1.022836  1.458759  2.287024

$c
[1] "a" "b" "c" "d" "e" "f"

$d
$d[[1]]
function (x)  .Primitive("sin")

$d[[2]]
function (x)  .Primitive("cos")
```

```
$d[[3]]
function (x, base = exp(1))  .Primitive("log")

$d[[4]]
function (x)  .Primitive("exp")
```

講義資料

授業で用いたスクリプトは以下のリンクに随時掲載する．
http://www.is.titech.ac.jp/ s-taiji/lecture/dataanalysis/dataanalysis.html

また，レポート問題や講義資料は OCW にもアップロードする．

R 関連参考資料

- CRAN (Complete R Archive Network)．R 本体および千を越す膨大な貢献パッケージが入手できる．日本のミラーサイトは http://cran.md.tsukuba.ac.jp
- RjpWiki．日本の R ユーザーが運営する情報サイト．各種 Tips や質問コーナーがある．URL は http://www.okada.jp.org/RWiki/
- 岡田昌史編「The R Book データ解析環境 R の活用事例集」，九天社 (2004)
- 間瀬茂著「R プログラミングマニュアル」，数理工学社 (2007)