



# 総務省 ICTスキル総合習得教材

【概要版】eラーニング用 

【コース4】 オープンデータ・ビッグデータ利活用事例 

## 4-3：プログラミングによるビッグデータの分析（R）

	1	2	3	4	5
【コース1】 データ収集					
【コース2】 データ蓄積					
【コース3】 データ分析					
【コース4】 データ利活用			◆		

# 本講座の学習内容（4-3：プログラミングによるビッグデータの分析（R））

## 【講座概要】

- 統計分析ソフトウェアRとRstudioの概要を示し、ダウンロード・インストール方法を紹介します。
- Rstudioの画面構成と基本操作を説明します。
- Rstudioを用いたExcelファイルの読み込み方法、回帰分析の実行方法を説明します。
- Rを用いることで高度な分析、大容量ビッグデータの分析ができることを示します。

## 【講座構成】

実 習 紹 介	
	[1]RとRstudioのダウンロード・インストール
	[2]RとRstudioの基本操作
	[3]Rstudioにおけるデータ分析

## 【学習のゴール】

- ✓ 統計分析ソフトウェアRとRstudioの概要を把握します。
- ✓ Rstudioにおける画面構成、基本操作を把握し、プログラミングの具体例を理解します。
- ✓ Rを用いることで高度な分析、大容量ビッグデータの分析ができることを把握します。

# 統計分析ソフトウェアRとRstudio

◆「R」は無料で利用できる統計分析用ソフトウェア（プログラミング言語）、  
「R studio」は「R」を快適に利用するための統合開発環境です。

- この講座では、ビッグデータ分析をはじめとする様々な分析に活用できるR（アール）を説明します。
  - Rは、Windows、Macintosh、Linuxにインストールできる無料のソフトウェアであるとともにプログラミング言語です。
- Rは、データ分析に特化した言語で、データ分析の初心者から専門家まで幅広い人気があります。
  - 様々なソフトウェアの制作に利用されるC言語やJavaといった汎用プログラミング言語と異なり、Rはデータ分析がしやすい設計になっています。
  - 米国電気電子学会が人気のあるプログラミング言語を示した「The Top Programming Languages 2017」において、Rは第6位になっています。
- Rstudioは、Rを快適に利用することができる統合開発環境です。
  - 統合開発環境（IDE: Integrated Development Environment）は、一つのソフトウェアの中に入力欄、出力欄、データ欄等が統合されて表示されることで、プログラミング等による開発を行いやすくする環境です。
  - R studioは、無料で利用できるオープンソース版と優先的なサポートが受けられる商用ライセンスがあります。

統計分析ソフトウェアRのロゴ



© 2016 The R Foundation.

Rは  
第6位  
の人気

The Top Programming Languages 2017の上位10位

Language Rank	Types	Spectrum Ranking
1. Python	🌐 🖥️	100.0
2. C	📱 🖥️ 🖨️	99.7
3. Java	🌐 📱 🖥️	99.4
4. C++	📱 🖥️ 🖨️	97.2
5. C#	🌐 📱 🖥️	88.6
6. R	🖥️	88.1
7. JavaScript	🌐 📱	85.5
8. PHP	🌐	81.4
9. Go	🌐 🖥️	76.1
10. Swift	📱 🖥️	75.3

【出典】米国電気電子学会（IEEE）

<https://spectrum.ieee.org/static/interactive-the-top-programming-languages-2017>

統合開発環境R studioのロゴ



RStudio is trademarks of RStudio, Inc

# RとRstudioのダウンロード

## ◆RとR studioは、誰でもウェブサイトからダウンロードすることができます。

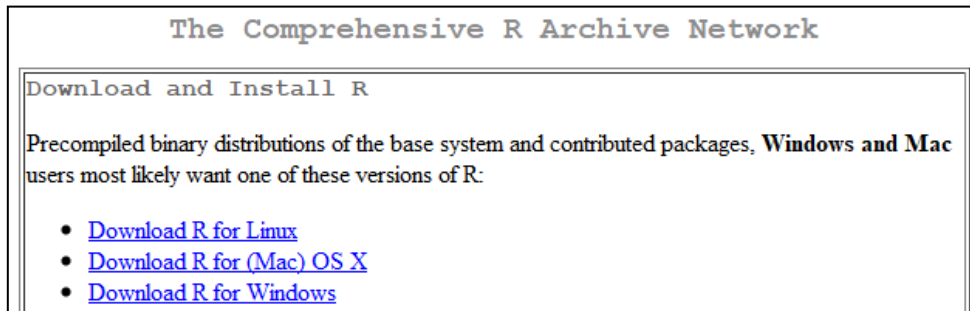
- Rのインストール用ファイルは、CRANに参加する統計数理研究所のウェブサイトからダウンロードすることができます。 <https://cran.ism.ac.jp/>

What are R and CRAN?



- CRAN (Comprehensive R Archive Network) は、Rに関するファイルを蓄積・提供する国際ネットワークです。
- 2017年10月時点における上記URLのウェブサイトの表記は概ね英語ですが、英単語が分かれば、ダウンロードやインストールに支障はありません。
- OSへインストールするためのRには、Windows版、Macintosh版、Linux版がありますが、この講座ではWindows版で説明します。
- Windowsを利用している場合は「Download R for Windows」をクリックした後、「base」も文字をクリックした後に表示されるWindows版のダウンロードボタンをクリックして下さい。

### OSに応じたRの選択画面



### Windows用Rのダウンロード画面

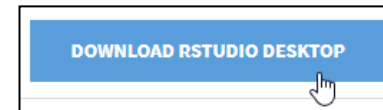


- Rstudioのインストール用ファイルは、Rstudioのウェブサイトからダウンロードできます。

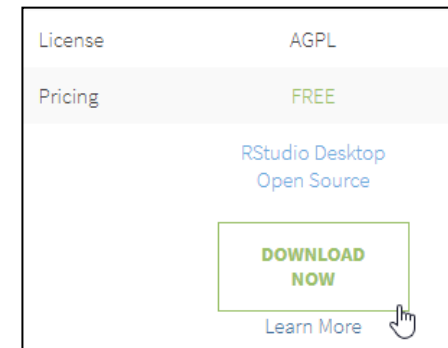
<https://www.rstudio.com/products/rstudio/download/>

- Rstudioには、各PCの中のRを実行するデスクトップ版と離れたサーバ上のRを実行するサーバ版がありますが、一般にはデスクトップ版を利用します。
- Rstudioのトップページからの移動する場合は、まず画面上部の「Products > Rstudio」を選択してください。次に表示される画面で「Open Source Edition」の欄にある「DOWNLOAD RSTUDIO DESKTOP」のボタンを押します。続いて表示される画面でオープンソース版の「DOWNLOAD NOW」をクリックします。

### デスクトップ版のダウンロードへのリンク



### オープンソース版のダウンロードボタン



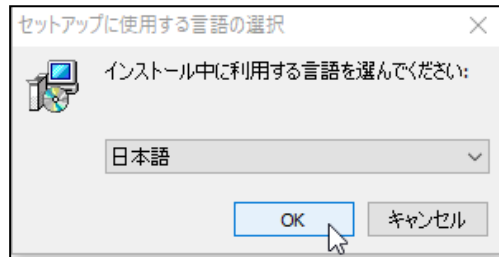
# RとRstudioのインストール

## ◆RとR studioは、マウスのクリックだけで簡単にインストールすることができます。

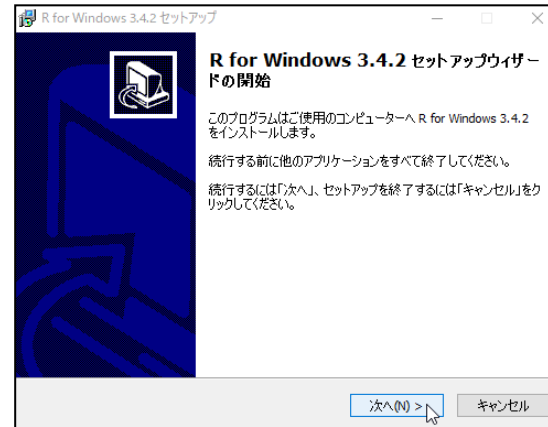
### ●Rのインストールにおいては、全て初期設定で「OK」や「次へ」で進めて、問題ありません。

- 設定内容が把握でき、変更したい方は、インストール先のフォルダの指定、32bit版か64bit版等の選択をして下さい。設定内容が把握できない方や細かい設定を気にしない方は、全て初期設定でのインストール、32bit版と64bit版の両方のインストールで構いません。

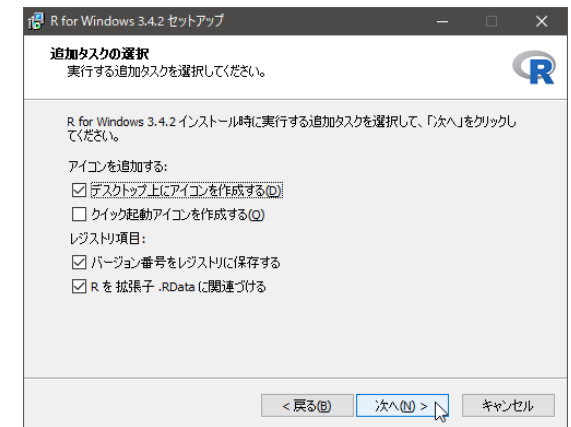
### Rのインストールの言語選択



### Rのインストール開始画面



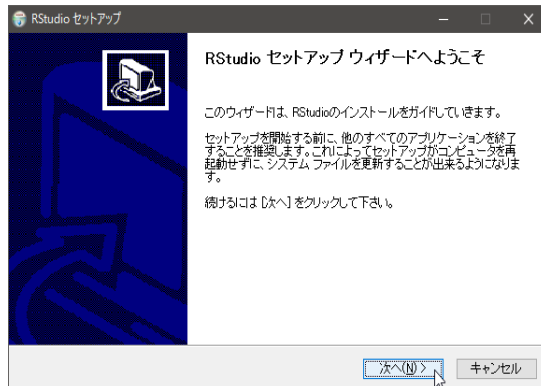
### Rのインストール時の最後の選択



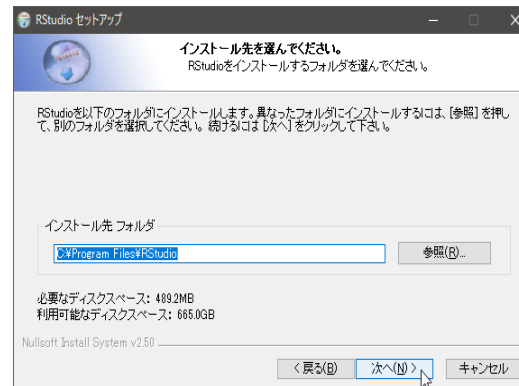
### ●Rstudioのインストールも、全て初期設定で「次へ」で進めて、問題ありません。

- 初期設定でインストールを完了すると、スタートメニューの中にRstudioのショートカットができます。これをクリックすると、Rstudioが起動します。

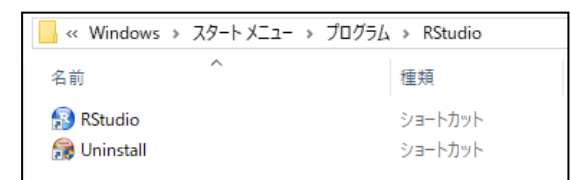
### Rstudioインストール開始



### Rstudioのインストール先指定



### プログラムフォルダ内のショートカット



# Rの起動と基本操作

## ◆Rを直接操作して、プログラミングと出力の関係を確認します。

### ●RおよびRstudioのインストール後は、右下のようなショートカットアイコンが表示されます。

- 「R i386」は32ビット版のRを指し、「R x64」は64ビット版のRを指します。Windowsの場合は、利用しているWindowsが32ビット版なら「R i386」、64ビット版なら「R x64」を使って下さい。利用しているWindowsが32ビットか64ビットか分からない場合は、どちらでもプログラムが動く「R i386」を使って下さい。



### ●まず、Rを直接操作するためにRのショートカットアイコンをクリックして起動します。

- Rの基本部分は日本語化がされており、初期画面にはRのライセンスに関する日本語での説明が表示されます。

### ●Rの直接操作、プログラミング体験として、中央下の枠内の黒字の部分の入力し、出力を見ます。

- Rでは「#（番号記号、ナンバーサイン、ハッシュ）」の右側をプログラミングとしての読み込み時に無視します。「#」の右側には日本語でも説明書きやコメントを書くことができます。

#### Rの初期画面の表示

```

RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ

R Console

R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-mingw32/x64 (64-bit)

R は、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、'license()' あるいは 'licence()' と入力してください。

R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で引用する際の形式については
'citation()' と入力してください。

'demo()' と入力すればデモをみることができます。
'help()' とすればオンラインヘルプが出ます。
'help.start()' で HTML ブラウザによるヘルプがみられます。
'q()' と入力すれば R を終了します。

> |

```

#### Rへのプログラムコード入力①

```

#足し算としての「1+2」
1+2
#Rで変数を作る場合は
#「変数名 <- 変数の中身」で入力
# x に10、yに20を入力
x<- 10
y<- 20
#xとyの足し算としての z
z=x+y
#変数名を入力すると、変数の値を出力
z
#全体を()でくくると、計算と同時に出力
(zz=x*y)

```

#### Rの出力

```

> #足し算としての「1+2」
> 1+2
[1] 3
> #Rで変数を作る場合は
> #「変数名 <- 変数の中身」で入力
> #xに10、yに20を入力
> x<- 10
> y<- 20
> #xとyの足し算としてのz
> z=x+y
> #変数名を入力すると、変数の値を出力
> z
[1] 30
> #全体を()でくくると、計算と同時に出力
> (zz=x*y)
[1] 200

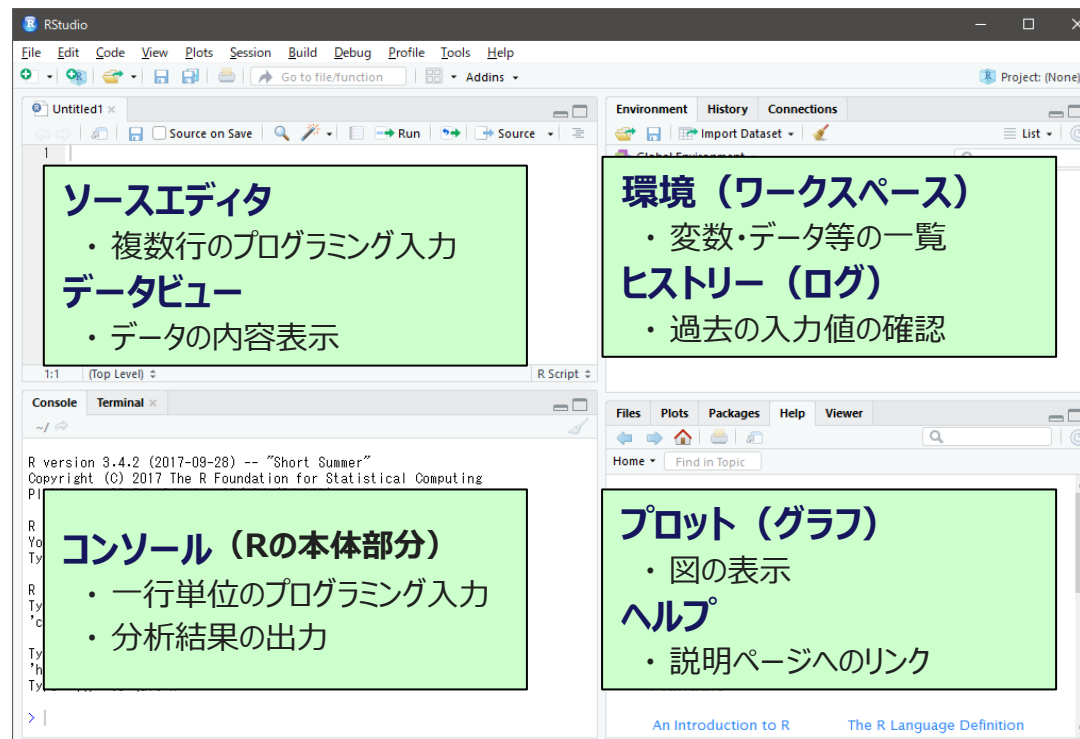
```

# Rstudioの画面構成

## ◆ Rstudioは分割した画面構成によって、Rのプログラミングを効率的に行えます

- Rstudioのショートカットアイコンをクリックすると、分割された画面構成のRstudioが起動します。
  - Rstudioには、公式の日本語版はありませんが、初歩的な英単語の知識で概ね読めることに加えて、ウェブ上の無料翻訳サービスを活用すれば、英語が苦手でもRstudioの利用に支障はありません。
  - 初期状態で画面の左側が縦に分割されていないのは、画面上側のメニューの左端にある[File]→[New File] →[R Script] を選択します。
- Rstudio内では分割された各パネルで、入力欄・出力・データー一覧・グラフと機能分化しています。
  - Rstudioでは分割された各パネルにタブ（つまみボタン）が付いており、パネル内の表示内容や表示対象を変えることができます。
  - Rstudioの画面構成は、メニューの[Tools]→[Global Options] →[Panel Layout]から、利用者の好みに合うようにカスタマイズできます。

### 初期設定におけるRstudioの画面構成（主なタブの内容）



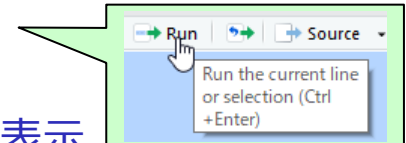


# Rstudioへの入力と画面出力

◆ Rstudioを使うと、変数データ一覧やグラフを確認しながら、プログラミングができます。

● 右下の画像では、ベクトル形式のデータを操作、線付きの散布図（グラフ）の描画を行っています。

- Rstudioでは、ソースエディタからプログラムコードを実行したい範囲を選択後、「Run」のボタンをクリックしてください。
- ベクトルは、数値を横（行）または列（縦）に並べたものを指し、数値を束ねたもののイメージです。



## Rへのプログラムコード入力②

### 2種類のベクトルの記入

```
v1<- c(1, 2, 3, 2, 1)
```

```
v2<- c(10, 20, 30, 40, 50)
```

#ベクトル同士の足し算（表示付）

```
(plus_v1v2=v1+v2)
```

#2つのベクトルを横に並べて行列作成（表示付）

```
(set_v1v2=cbind(v1, v2))
```

### 統計関数の利用

#平均値mean

```
mean(plus_v1v2)
```

#基本統計量セットsummary

```
summary(plus_v1v2)
```

#「set\_v1v2」を線付きで散布図で青で表示

```
plot(set_v1v2 ,type="o", col="blue")
```

## Rstudioの4分割画面の表示

**ソースエディタ 入力**

**データビュー データの表示**

	v1	v2
1	1	10
2	2	20
3	3	30
4	2	40
5	1	50

**環境（ワークスペース） 変数・データ一覧**

Object	Class	Attributes	Values
set_v1v2	num	[1:5, 1:2]	1 2 3 2 1 10 20 30 40 50
plus_v1v2	num	[1:5]	11 22 33 42 51
v1	num	[1:5]	1 2 3 2 1
v2	num	[1:5]	10 20 30 40 50

**コンソール 結果出力**

```
> #2つのベクトルを横に並べる (表示付)
> (set_v1v2=cbind(v1, v2))
  v1 v2
[1,] 1 10
[2,] 2 20
[3,] 3 30
[4,] 2 40
[5,] 1 50
> ###統計関数の利用
> #平均値mean
> mean(plus_v1v2)
[1] 31.8
> #基本統計量セットsummary
> summary(plus_v1v2)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 11.0   22.0   33.0   31.8   42.0   51.0
> #5行2列の行列を線付きで散布図で表示
> plot(set_v1v2 ,type="o", col="blue")
```

**プロット（グラフ） グラフ出力**

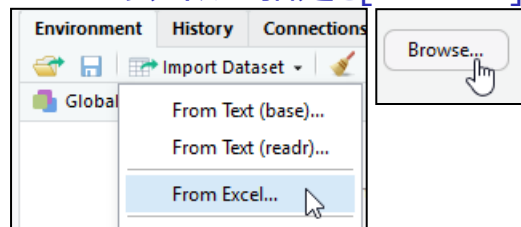


# RstudioにおけるExcelファイルの読み込み

## ◆Rstudioでは、簡単にExcelデータを読み込むことができます。

- Rstudioの標準設定における右上のパネルの[Import Dataset]から外部のデータを読み込みます。
- Excelファイルを取り込む場合は[From Excel]→[Browse]とクリックし、データの入ったExcelファイルの選択後、プレビューでデータの内容を確認してから [Import]をクリックします。
  - Rstudioの標準設定とするフォルダは、[Tools]→[Global Options]→[General] にある「Default working directory」から変更できます。
  - Excelファイル内の分析用データは1行目に変数名、2行目以降に一行ずつ個別の標本のデータが入っている形式にしておきます。
  - Rで日本語のファイル名を取り込む設定もありますが、半角英数字のファイル名にしておくと、データ読み込み時のエラーの心配がありません。
  - Excelファイルの中の各セルに入っているデータは、文字データの列であれば日本語が含まれていても問題ありません。
  - データがプレビューに表示されている状態では、[Code Preview] にデータと読み込みに対応するプログラムコードが表示されます。このコードをコピーして、ソースエディタに貼り付けることで、次回以降の同じデータ読み込みをする際にプログラムコード内で行えます。

### Excelファイルの指定と[Browse]



### プレビューによるデータ内容の確認

Import Excel Data			
File/Url:			
C:/data_ols.xlsx			
Data Preview:			
通し番号 (character)	y (double)	xα (double)	xβ (double)
標本1	109.33219	4.423869	17.69999
標本2	112.34647	3.743842	16.95618

### プログラムコードと[Import]



### Rstudio内に取り込んだExcelデータの表示

RStudio Environment Panel				
Environment History Connections				
Global Environment				
Data				
data_ols 50 obs. of 4 variables				
通し番号: chr "標本1" "標本2" "標本3" "標本4" ...				
y : num 109.3 112.3 164.5 94.3 127.6 ...				

RStudio Data Viewer				
	Filter			
通し番号	y	xα	xβ	
1 標本1	109.33219	4.423869	17.69999	
2 標本2	112.34647	3.743842	16.95618	
3 標本3	164.51316	16.626036	18.54085	

# Rstudioにおける回帰分析

## ◆Rでは読み込み済のデータに対して、1行のプログラムで回帰分析が実行できます。

- Rstudioで読み込んだExcelファイルはデータフレームと呼ばれる形式となり、データフレーム形式の中の各列は、「データフレーム名\$列名（変数名）」で指定することができます。
  - 標準的な読み込み設定では、Excelファイル上のデータの1行目が列名（変数名）となります。
  - データフレームの中の列名（変数名を）変更したい場合は「`names(data_ols) <- c("新列名1","新列名2","新列名3")`」と順に指定したり、「`names(data_ols)[3]<-"新列名3"`」と列の番号を指定して、変数名を改めることができます。
- Rにおける回帰分析は、「`lm(被説明変数 ~ 説明変数1 + 説明変数2 +...)`」という1行のプログラムコードで実施できます。
  - 前のスライドで取り込んだExcelデータに関する回帰分析の結果を「`lm_result`」という名前のデータ(リスト形式)として保存する場合は、「`lm_result<-lm(data_ols$y ~ data_ols$xα + data_ols$xβ)`」と入力します。

### ソースエディタにおける回帰分析のプログラムコード入力

```
#data_ols内のyを被説明変数、xαとxβを説明変数として回帰分析
lm_result<-lm(data_ols$y ~ data_ols$xα + data_ols$xβ)
```

## ●回帰分析の結果がデータ（リスト形式）をクリックして、データビューに分析結果が表示されます。

- 標準的な読み込み設定では、Excelファイル上のデータの1行目が列名（変数名）となります。

### 環境（ワークスペース）における表示

Data	
data_ols	50 obs. of 4 variables
lm_result	List of 12

### データビューに表示される回帰分析の結果の内容

Name	Type	Value
lm_result	list [12] (S3: lm)	List of length 12
coefficients	double [3]	21.96 3.20 4.61
residuals	double [50]	-8.382 0.239 3.841 -3.592 5.016 -5.207 ...

# Rにおけるパッケージの利用・Rにおけるビッグデータの活用

## ◆Rはパッケージを利用することで様々な出力、高度な分析を簡単に実行できます。

### ●複数の回帰分析の結果を並べて表示して、比較したい場合には、「memisc」パッケージが便利です。

- Rにおいて、Rのプログラムコードを配布用にとりまとめたものを「パッケージ」と言います。
- インターネット上のCRANに保存されているパッケージを初めて使う場合は、プログラムコードに「`install.packages("パッケージ名")`」と入力し、PC内にパッケージをダウンロード・インストールしてください。（一度、PCにインストールすれば2回目以降のプログラムコードへの記載は不要です。）
- PC内にインストールされているパッケージは、プログラムコードに「`library(パッケージ名)`」と入力した後に使うことができます。

### ソースエディタにおけるプログラムコード入力

```
#xα、xβのそれぞれ1変数で単回帰して結果を格納
lm_res2<-lm(data_ols$y ~ data_ols$xα)
lm_res3<-lm(data_ols$y ~ data_ols$xβ)
```

```
#パッケージ「memisc」のインストールと利用宣言
install.packages("memisc")
library(memisc)
```

```
#パッケージmemisc内のmtable関数を利用
#3つの回帰分析の結果を並べて表示
mtable(lm_result, lm_res2, lm_res3)
```

### mtable関数(memiscパッケージ)の出力

	lm_result	lm_res2	lm_res3
(Intercept)	21.962* (9.353)	103.421*** (4.522)	3.077 (20.695)
data_ols\$xα	3.203*** (0.231)	4.051*** (0.347)	
data_ols\$xβ	4.609*** (0.506)		7.449*** (1.035)
R-squared	0.906	0.740	0.519
adj. R-squared	0.902	0.734	0.509
sigma	8.612	14.174	19.260
F	226.088	136.289	51.811
p	0.000	0.000	0.000
Log-likelihood	-177.059	-202.497	-217.828
Deviance	3486.040	9643.413	17805.419
AIC	362.119	410.994	441.656
BIC	369.767	416.730	447.392
N	50	50	50

回帰分析の結果表示において、Interceptは切片の高さを表し、説明変数の値が全て0の場合における被説明変数の予測値に対応します。

効果がありそうな説明変数には「\*」を付けることが、分析結果の表記において慣例となっています。

「R-squared」は決定係数を意味し、0以上1以下の値をとる回帰分析の当てはまり度合いの指標です。

### ●Rのパッケージを使うと、機械学習の高度な分析も簡単なプログラムコードで実行できます。

- 本格的な機械学習のデータ処理・分析には、Python（パイソン）というプログラミング言語が優れており、人気があります。

### ●64ビット版のRを使うと、大容量のビッグデータのデータ処理・分析が可能です。

- 64ビット版のRでは、メインメモリの容量を上限としてデータを格納することができ、GB（ギガバイト）単位のデータ処理・分析が可能です。