フリーソフトR言語およびその応用

ザン　ピン*，　高田　正之**

要　約

R言語で統計分析を検討しはじめる方のために，Rの初歩から実行結果までコンパクトにまとめた。Rの起動画面，データファイルの読み込み，身近な例を通じて，統計の回帰，検定，および因子分析に限定し，実行とその結果を提示した。まだ初期段階だが，Rを利用し，文系学生のモデリングの模索も試みた。

キーワード：R言語，回帰，検定，因子分析，PBL

1. はじめに

フリーの統計ソフトはさまざまな道のりを経て，ようやくRという形にたどり着いた。今，その知名度は統計有償ソフトSPSS，SASに劣らないほどになり，信頼性も高まり，教育，研究，企業で幅広く応用されている。

時代性から，Rは伝統的な統計学より，データマイニング，データサイエンスのツールとしてよく紹介されている。また，SASを利用するユーザーには馴染みやすいが，PCのSPSSユーザーにとって，Rのスクリプト言語の性質から，やはりやや敷居が高い。また，著者らの経験から，他の研究に専念していて，短時間内に利用できるツールを探している研究者に，あるいは，学生らに好奇心が保っているうちにまずRに何かができるかを見せるのは好ましい。そういう考えで，本稿は統計学基礎の回帰，検定，および因子分析に絞り，Rを素早く試し，イメージを掴めるためにコンパクトにまとめる。そしてさらに，Rについての文系大学の授業での学生の反応について考察した。

2015年11月30日受付

* 江戸川大学 情報文化学科准教授　数理計画
** 江戸川大学 情報文化学科教授　ソフトウェア

2. Rによるデータの作成

Rのダウンロードは簡単で，短時間でできる。起動の初期画面は図1の通り。

例えば，6人の身長のデータ（150, 155, 160, 165, 170, 175）をxに一時入れ，その基本統計量を示すのは以下のように入力する[(1)]（最後は改行すること，以下同様）。

```
x <- c(150, 155, 160, 165, 170, 175)
summary(x)
```

出力，つまり，実行結果は以下になる（一部）。

```
Min. 1st Qu.  Median    Mean 3rd Q
 150     160     165     164     1
```

現在，データはExcelで作成されることは多い。例えば，図2のような，性別，身長と体重データのファイルht_wt.csv[(2)]を読み込むには[(3)]，こう書く。

```
x <- read.csv("ht_wt.csv")
```

また，読み込まれたデータxから身長の列，条件は女性かつidが20以下のデータだけを取り出すにはこう書く。

```
y <- x$ht
x_f20 <- subset(x,sex=="f" & id<21)
```

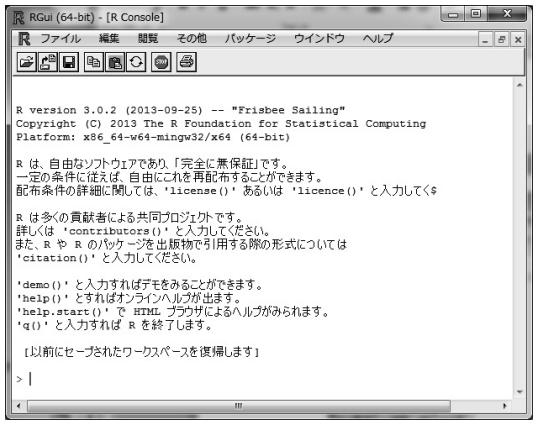

図 1　起動画面

	A	B	C	D
1	id	sex	ht	wt
2	1	f	159.1	58.1
3	2	f	145.9	49
4	3	f	154.8	50.2
5	4	f	147.2	47.3
6	5	f	162.2	79.5
7	6	f	157.6	61.5
8	7	f	156.3	59.6
9	8	f	161	64.7
10	9	f	152.6	56
11	10	f	158.9	57.2
12	11	m	171	83.3

図 2　EXCEL ファイル形式

y や x_f20　だけでデータを確認できる。

身長のヒストグラムを描くには(図3)こう書く。

```
hist(y)
hist(wt,breaks=30)
```

関数によるデータの生成もできる。例えば，0 から 50 までの整数を生成して x1 に代入し，それを横軸に，縦軸は平均 25，分散 10 の正規分布の密度関数をグラフで表示する場合，以下の通り入力する（図 4）

```
x1 <- 0:50
plot(x1, dnorm(x1,25,10))
```

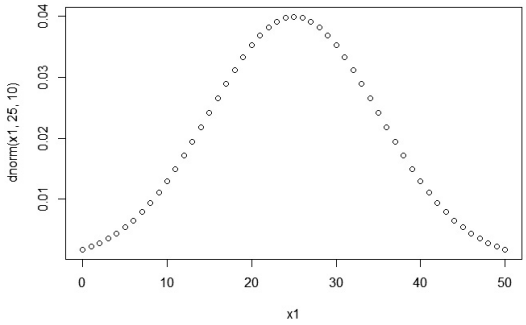

図 4　正規分布図

3. 回帰，検定，因子分析

3.1　回帰

先ず，x に格納されている身長と体重のデータ散布図を

```
plot(x$ht,x$wt)
```

で描く。さらに，散布図の上に直線（線形モデル）を描くには以下のように入力する。

```
res <- lm(x$wt~x$ht, x)  (4)
abline(res)
```

実行結果は図 5 の通りである。　相関と他の係数を求めたいとき，

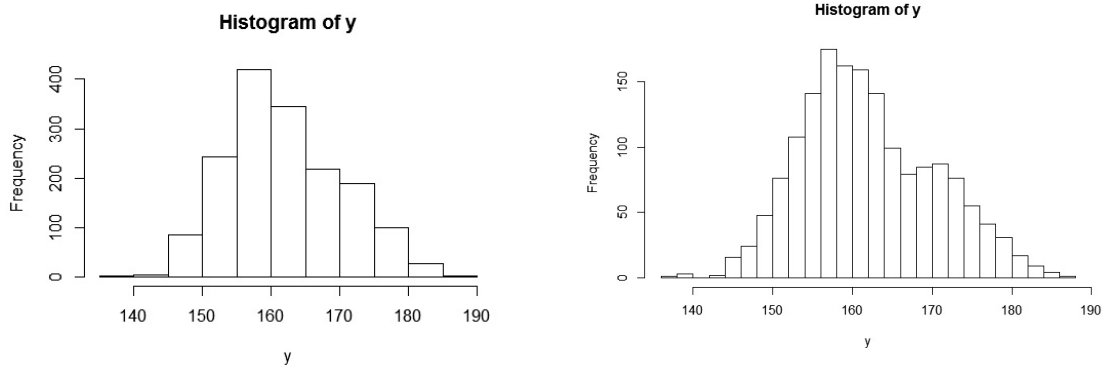

図 3　ヒストグラム

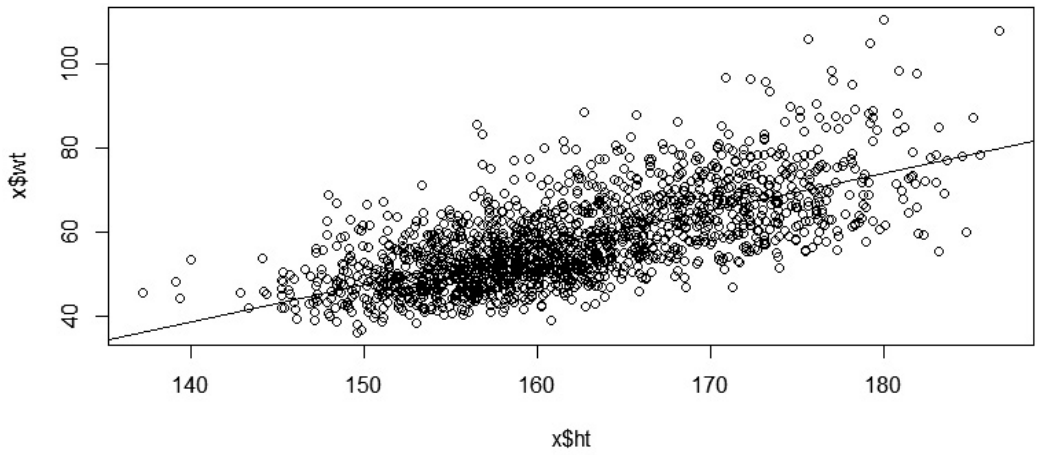

図 5　散布図と回帰直線

```
cor(x$ht,x$wt)
summary(res)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -84.81831    3.79898  -22.33   <2e-16 ***
x$ht          0.88218    0.02342   37.66   <2e-16 ***
---
```

図 6　summary（res）出力の一部

図 6 のデータから，体重（キロ）＝－ 84.818+0.882 ×身長（センチ）という関係式になる。

重回帰の場合（fat の要素を入れて）

```
cor(x$wt,x$ht+x$fat)
res <- lm(x$wt~x$ht+x$fat, x)
```

3.2　t 検定

例えば，男女の身長差を t 検定で行う。先ず，男女それぞれの標準偏差を sd で確認する。sd(x_m$ht）と sd（x_f$ht）の値はそれぞれ，5.94 と 5.22，近似的に等分散（T，すなわち，True）と仮定する。t 検定は以下のコマンドで，実行結果は図 7 である。

```
t.test(x_m$ht,x_f$ht,var.equal=T)
```

t 値は 46.41，95％の信頼区間の外にある（P 値はかなり小さい数値）。帰無仮説：「男女の平均の身長差が 0」を棄却。仮説は有意である（当然な結果）。

3.3　因子分析

図 8 のデータ（[3]）を見てみよう。ファイルを読み込み（header=F と指定し，数値データをヘッダとして扱うことを防ぐ), 2 要素に因子分析，実行結果は図 9 に示す。

```
        Two Sample t-test

data:  x_m$ht and x_f$ht
t = 46.4139, df = 1638, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 12.52156 13.62656
sample estimates:
mean of x mean of y
 170.2369  157.1628
```

図 7　t 検定の結果

89	90	67	46	50
57	70	80	85	90
80	90	35	40	50
40	60	50	45	55
78	85	45	55	60
55	65	80	75	85
90	85	88	92	95

(a) seiseki2.csv

	算数	理科	国語	英語	社会
A	89	90	67	46	50
B	57	70	80	85	90
C	80	90	35	40	50
D	40	60	50	45	55
E	78	85	45	55	60
F	55	65	80	75	85
G	90	85	88	92	95

(b) seiseki.csv

図 8　成績ファイル

```
ss<-read.csv("seiseki2.csv", header=F)
ss.fac<-factanal(ss,factors=2,sc="regression")
ss.fac
```

```
Loadings:
   Factor1 Factor2
V1           0.997
V2 -0.188    0.967
V3  0.871
V4  0.997
V5  0.989  -0.128

                Factor1 Factor2
SS loadings       2.768   1.946
Proportion Var    0.554   0.389
Cumulative Var    0.554   0.943
```

図 9　因子分析の結果

因子 2（Factor2）のところは，算数，理科，つまり，理系の因子負荷量（0.997, 0.967）が高い，因子 1 には，文系の 3 科目，国語，英語，社会の因子負荷量が高い。因子 2 の累積寄与率は 0.943 と高い。因子得点を表示するには，さらに次のように，

```
ss.fac$sc
```

を実行する必要がある。その結果は図 10 に示される。

図 8 と図 10 を比較し，個体 A は因子 2 の得点が高く，因子 1 の得点が低い，つまり，理系が強い。個体 B はその逆。最後の個体 G は 2 因子の得点が高い，理系文系関係なく成績がよい。

結果を図示化するには，例えば，以下のように入力する（図 11）。

```
          Factor1    Factor2
[1,] -0.7872337  1.0078371
[2,]  1.0013394 -0.6725305
[3,] -0.9919085  0.5958555
[4,] -0.8939458 -1.5278910
[5,] -0.3836573  0.4519010
[6,]  0.6283699 -0.8225840
[7,]  1.4270361  0.9674119
```

図 10　因子得点

```
colnames (ss) <-c (" 算数 "," 理科 "," 国語 "," 英語 "," 社会 ")
barplot(ss.fac$loading [,1])
barplot(ss.fac$loading [,2])
```

4. R による PBL

4.1 ディスカッション

PBL というのは Project-Based Learning 課題解決型学習である。

今回は一般授業で一部の時間を利用し，回帰分析について紹介・演習・課題を実施した。

文系の学生は，数式を使い，予測するという経験が少ない。そのため，できるだけ身近な問題を取り上げた。学生全員が 2-3 節に書かれてある身長と体重のファイルを読み込み，散布図と回帰式を実行した。学生がまず疑問を感じるのは，切片の大きなマイナスの値である。それに対し，軸を伸ばして描く。さらに，各自が身長を代入し，体重を計算する。それらを通じて，回帰式を理解・実感してもらう。

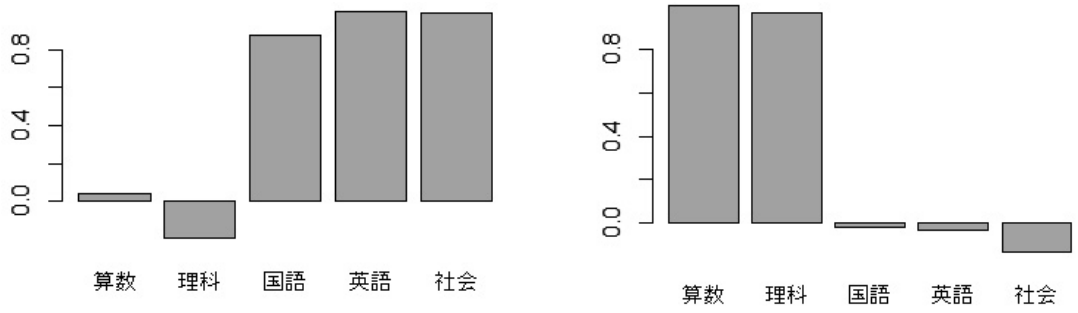

図 11　因子得点棒グラフ

次に，学生が回帰モデルを提案するという課題に移る。身近な問題として，「明日の気温を予測する」という例を通じて，全員がアイデアを出し，ディスカッションしてもらう。

しばらく沈黙した後，「課題の意味を理解してない，我々は何をすればよいのか」とある学生から逆に問いかけられた。そこで，「明日の気温は前例の体重，身長の位置にベターと思われる要素を考える」と説明し，議論が一気に活発になった。

湿度，気圧，晴れ，日差し，経度，今月の気温，去年の同じ日の気温，…

やはり，未経験のため，ローカルな要因，あるいは，通常の天気予報に影響され，シンプルかつ有効な数式モデルに不可欠なベター要素とその表現に直ぐにたどり着かない。

学生たちの議論後，季節を表現する平均気温や現在の気温を補足した。

ソフトによる演習・ディスカッション・思考を組み合わせて，短時間だが，学生にとっても，教員にとっても，有意義な授業を展開できた。

4.2　モデリング

次に，気象庁が公開している過去データ（図 12）に基づいてモデリングする。

「項目を 5 つ以内に制限し，もっとも精度の高い明日の気温を予測する」。質問・課題を明確にするために，前回のディスカッションの結果と合わせて，

明日の気温 <= （例年の）平均気温, . . .

とホワイトボードに書いて，印が付けられている気温について，「もっと限定したほうがよい」と付け加えた。

各自が紙に書く前に，「地点」，関連の「項目」と「期間」を選択し，CSV ファイルをダウンロードし，その場で見せた（図 13）。

一部の学生はモデルについてよく理解していないが，いくつか良い答案を出している学生もいた。

明日の気温　<=　平均気温，前日の気温，湿度，気圧，風向き

明日の気温　<=　平均気温，最高気温，最低気温，日照時間

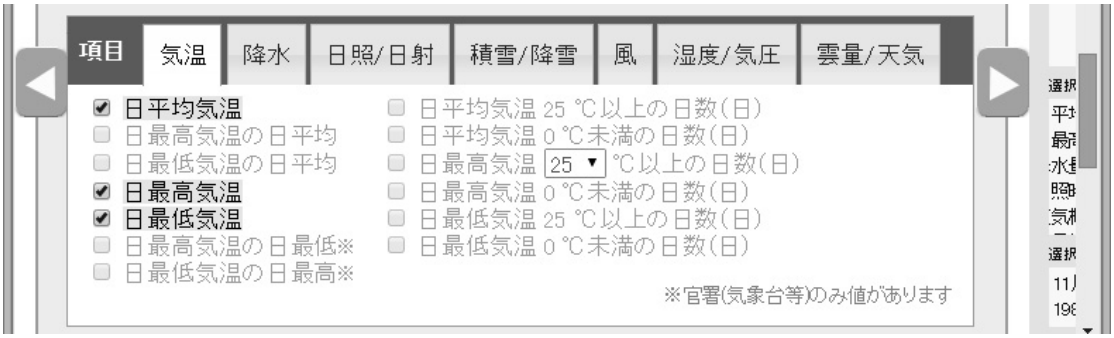

図 12　気象庁の過去データ（http://www.data.jma.go.jp/gmd/risk/obsdl/）

D18 fx 12.2

	A	B	C	D	E	F	G	H	I
1	ダウンロードした時刻:20	我孫子							
2	年月日	平均気温(°	最高気温(°	最低気温(°	降水量の合	日照時間(時	最大風速(m	最大風速(m	平均風速(m/s
3	2015/10/15	16.2	22.6	10.8	0	6.4	3.9	東	1.5
4	2015/10/16	15.4	16.8	14.4	7	0	2.6	北	1.7
5	2015/10/17	17	20.2	14.9	1.5	0.1	2.5	北北東	1.5
6	2015/10/18	17.7	22.9	13.4	0	5.9	2.8	東	1.4
7	2015/10/19	16.6	23.1	12	0	10.2	3.8	東	1.3

図 13　気象庁の過去データ，修正されたダウンロードファイル

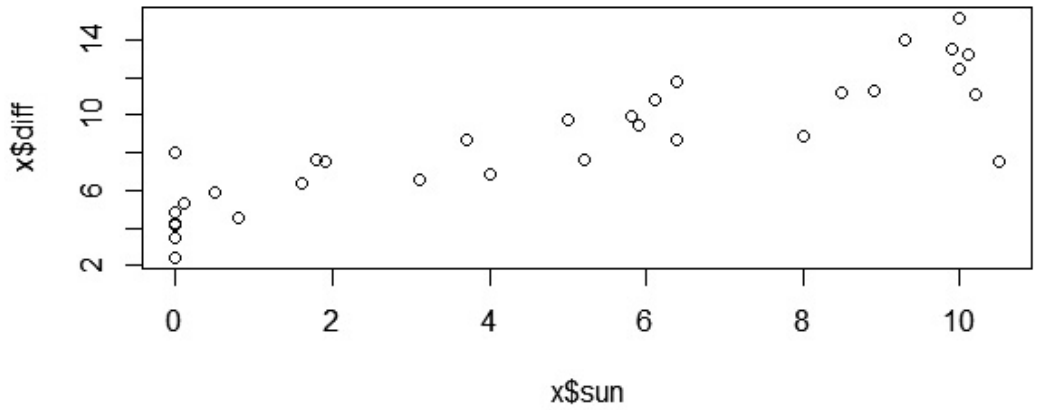

図 14　日照時間と気温差

明日の気温　<=　平均気温，降水量，日照時間

明日の気温　<=　平均気温, 最低気温, 降水量, 日照時間，平均風速

具体的にモデリングするには，さまざまな修正の繰り返しが不可欠。データの不備もあり，データの加工も時に必要である。上の例の場合，同じ月日の平均気温を算出し，データのフォーマットの整理なども含まれている。データマイニングによる発見も面白い。例えば, 素早く, [直近 1 ヵ月] (2015/10/15-2015/11/15) を選択し，気温差（最高気温―最低気温）を新たに列に加えて，日照時間との散布図は図 14 の通り，比較的に強い相関 0.867 を持っている。

4.3　まとめ

R 言語系のフリーソフトであり，敷居が高いと思われるが，30 分程度の 3 回の授業で，以上のようなことが実践できた。学生はこのプロセスを通じて，三つの力,「コミュニケーション力」,「考える力」,「基礎力」を，ある程度学び，感じることができた。

参考文献

[1] 南風原朝和，心理統計学の基礎，統合的な理解のために，有斐閣，2002
[2] 南風原朝和，続・心理統計学の基礎，統合的な理解を広げ深める，有斐閣，2014
[3] 金明哲，R によるデータサイエンス，森北出版，2007
[4] W.Richert，L.P.Coelho，斉藤康毅（訳），実践機械学習システム，オライリー・ジャパン，2014
[5] 片岡厳，データサイエンス養成読本，技術評論社，2013
[6] 片岡厳，データサイエンス養成読本，技術評論社，2015
[7] 統計ソフト R の使い方，https://sites.google.com/site/webtextofr/home

《注》

(1) ここに，注意すべき点は，c は小文字でなければならない。「<-」の代わりに「=」も使える。メモ帳から入力，編集した内容を R にペーストするのも構わない。
(2) 参考サイト［7］の demodata.csv の一部
(3) ファイルはドキュメントフォルダに保存するのが簡単

である，別のフォルダに保存する場合，「ファイル」タブの「ディレクトリの変更」をクリックし，フォルダを選択する。

(4) この wt と ht の順番 に注意 ，他と逆