Rによるアンケート調査について

小　村　賢　二

要約；この研究ノートはRとSPSSとのインターフェイスを利用し、Rを使ったアンケート調査の方法を提案する。RはODBCであるからプログラミングを介してRとSPSSとSASは相互に利用可能である。従来アンケート調査を行うときSPSSやSASが一般的に使われてきた。Rはデータ解析とグラフィックスにおいて最新の統計理論を組み込んだ数多くのパッケージが導入されてグローバルスタンダードとして使われている。Rを活用すればSPSSでは分析できないことが多くの点で可能である。Rはアンケート調査や市場調査でも大量データが解析でき、対応できることを示す。

キーワード；R2.14.1 , SPSS, Rcmdr, foreign, logistic Regression, odds , syntax

序　はじめに

Rはデータに関してODBC(Open Data Base Connectivity)であり各種のデータの形式EXCEL, TEXT,SPSS, SAS,MINITAB等のデータを読み込んで分析できる。新しいタイプのものである。最近のSPSSは(Version 18以後) Rとのインターフェイスを持ち、Rのプログラミング(オブジェクト)をSPSSのシンタックスエディターに貼り付けてRのオブジェクトを実行できるようになった。R(R2.14.1)のデータ解析は広範囲に渡りSPSSのできない多くの分析も可能である。記述統計から、多変量解析、生存分析や時系列解析等多分野に渡り利用でき、最近ではテキストマイニング等のパッケージも含め膨大なデータ解析のソフトである。これが世界の共通財産をして世界中の人々から支持されており、利用に関して無料であることは素晴らしい。データの操作、解析、グラフ表現の操作が容易であり

システムやプログラムはオープンソース(open-source)で公開されている。Rは Windows, Macintosh, Unix と Linux でも走る。現在 R の Version は 2.14.1 である。(平成 24 年 2 月 20 日現在)

§1　データの読み込みについて

データの読み込みについて標本数が十分大きいとき(n>1000)には EXCEL のシートに、データベースを作成し標本数が少ないときには、R-コンソールまたは R-コマンダーのデータエディターに読み込ませる。これはインターフェイスが EXCEL であればキイボードからのデータの読み込みが容易であり、多人数のアンケート調査であってもタスクを分担すれば比較的短時間に入力からデータのマージ(併合)までデータベースを作成することができる。講義・実習において多人数で分担するとき、データの入力は半角・英数の入力モードであることに注意する。EXCEL では入力モードについて区別しないが R ではひらがな入力と半角・英数入力のデータに区別がある。データを USB に保存するには R-コンソールで作業領域についてディレクトリーを変更しておく必要がある。今 n=1007 のデータの一部分をデータ①とする。アンケート調査データのデータセットを anketo2.csv とする。

Gender	age	job	relations	cm	effects	goodsbuy
2	40.00	6	2	1	2	2
2	17.00	5	4	1	3	2
2	25.00	2	3	2	3	2
1	48.00	1	2	1	3	2
2	23.00	7	4	1	3	2
2	19.00	5	3	1	3	2
2	51.00	6	1	2	3	2
1	31.00	1	2	2	3	2
2	30.00	6	2	1	2	2
1	21.00	5	3	2	3	2

次のデータ②はRで読み込んだもので、SPSS に変換してRcmdr を通してデータをインポート(import)して data view したものである。R ではオブジェクト名や変数名はローマ字または英語であることが必要である。EXCEL や SPSS では調査表の raw データについて尺度に変換したものもあり、job=5 は? とデータ①では確認しづらい。次はRのデータセット anketo2.rdat でその一部分をデータ②とする。

Gender	age	job	relations	cm	effect	goodsbuy
女	40	主婦	少し密着	よく見る	少し影響される	ない
女	17	学生	ほとんどしていない	よく見る	余り影響されない	ない
女	25	公務員	余り密着していない	余り見ない	余り影響されない	ない

男	48	会社員	少し密着	よく見る	余り影響されない	ない
女	23	その他	ほとんどしていない	よく見る	余り影響されない	ない
女	19	学生	余り密着していない	よく見る	余り影響されない	ない
女	51	主婦	大変密	着余り見ない	余り影響されない	ない
男	31	会社員	少し密着	余り見ない	余り影響されない	ない
女	30	主婦	少し密着	よく見る	少し影響される	ない
男	21	学生	余り密着していない	余り見ない	余り影響されない	ない

R コマンダー(Rcmdr)において読み込んだデータエディターは調査表のデータの内容が分りやすいことが分かる。R-コンソールで　EXCEL のデータセット anketo2.csv を次の
anketo2 <-　read.csv(" anketo2.csv" , header=TRUE)
で読み込み、パッケージの選択で　foreign　と　Rcmdr を選択する。R-コマンダーのメニューから anketo2.csv や anketo2.sav データをインポートによって読み込ませる。 読み込むデータセット名を Dataset002a とする。R に SPSS のファイル anketo2.sav をインポートするとスクリプトウィンドウで次が表示される。

```
Dataset002a <- read.spss("C:/Users/komura/Desktop/anketo2.sav",
use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)
```

データセットの表示でデータセット②が表示され、
出力ウィンドウで以下が表示される。

```
> Dataset002a <- read.spss("C:/Users/komura/Desktop/anketo2.sav",
+     use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)
> library(relimp, pos=4)
>showData(Dataset002a, placement='-20+200', font=getRcmdr('logFont'),
+     maxwidth=80, maxheight=30)
```

ここで EXCEL から SPSS へデータをコピー&貼り付けて SPSS でアンケート調査を分析するには文字型の変数には名義尺度を付ける。
変数ビューで　名前 gender 型　数値から文字列に変更し、値　1=男性　、2=女性
と値付ける。文字列の変数について同様にする。作成したら、anketo2.sav で保存。
anketo2.sav の分析はメニューから分析、要約、度数分布、クロス表で行う。
アンケート調査の標本数が少ないときには Rーコンソールでデータベースを作成。R でもデータの merge は可能である。データはベクトルのオブジェクト作成より、入力し易い scan() 関数を使う。文字型の変数、性別は
gender <- scan(" ",　what=　" ")
としてデータを順に読み込む。age <- scan() とし同様に入力し、goodsbuy<-　scan(" " , what= " ")　として、データを読み込ませて、次のデータフレームを作成する。
anketo2 <-　data.frame(性別＝gender ,年令=age,　,商品購入=goodsbuy)
と入力し print(anketo2)でデータセットが表示される。

§２　アンケート調査の多次元分割表について

Ｒの分析はR-コンソールとＲコマンダーで分析できる。要約、度数分布、クロス集計ができる。以下 SPSS のシンタックスを合わせて一部分を表示する。Ｒ－コマンダーでデータセットの要約は以下のオブジェクトを入力する。

summary(Dataset002a)

gender

男	529
女	478

job

会社員	211
公務員	47
自由業	5
教員	12
学生	597
主婦	108
その他	27

Goodsbuy（商品購入）

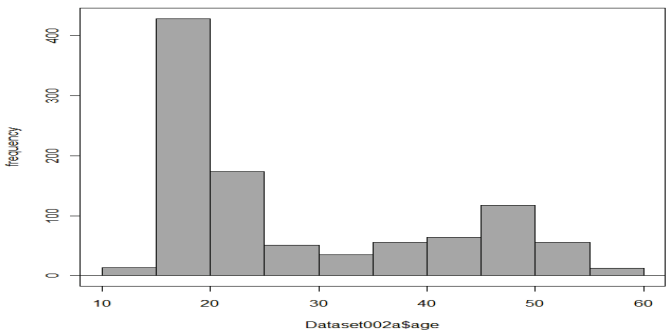
ある	131
ない	876

Ｒのヒストグラム作成のスクリプトと出力ウィンドウは以下である

Hist(Dataset002a$age, scale="frequency", breaks="Sturges", col="darkgray")

Hist(Dataset002a$age, scale="frequency", breaks="Sturges", col="darkgray")

ここで Sturges はスタージェスのルール；階級の個数ｋ　；$k = \log_2 n + 1$である。



SPSS について、プログラムの作成、即ちメニューの分析からシンタックスを貼り付ける。プログラムは実行手順に従えばシンタックスエディターに貼り付けて作成できる。

```
DATASET ACTIVATE データセット1.
SPSS のシンタックス
FREQUENCIES VARIABLES=gender job reations cm effect goodsbuy     ＃度数分布の作成
  /PIECHART FREQ
  /ORDER=ANALYSIS.
CROSSTABS                              ＃クロス集計
  /TABLES=gender BY relations
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.
CROSSTABS                              ＃クロス集計
  /TABLES=gender BY cm
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.
CROSSTABS                                ＃クロス集計
  /TABLES=gender BY effect
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.
CROSSTABS                          ＃クロス集計
  /TABLES=gender BY goodsbuy
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

SPSS では３重クロス（多重クロス表）が作成できないのでRを使う。
多元分割表はRと SAS のみ可能で行の変数に cm　列の変数に gender , 　コントロール変数　に　goodsbuy を入力する。アンケート調査や市場調査ではクロス表や多重クロス（分割表）は分析に大きなウエイトを持つ。度数分布表のみではデータの持つ構造的な解析ができない。以下はRのオブジェクトである。

```
library(abind, pos=4)
.Table <- xtabs(~cm+gender+goodsbuy, data=Dataset002a)
.Table
remove(.Table)
> library(abind, pos=4)
> .Table <- xtabs(~cm+gender+goodsbuy, data=Dataset002a)
```

> .Table

, , goodsbuy = 商品購入ある

	gender
cm(コマーシャル)	男　女
よく見る	52　49
余り見ない	0　12
コマーシャルは他に変える	10　8

, , goodsbuy =商品購入ない

	gender
cm(コマーシャル)	男　女
よく見る	288　278
余り見ない	85　54
コマーシャルは他に変える	94　77

以下はコントロール変数に cm を入れたときの多重分割表である。

> .Table <- xtabs(~gender+goodsbuy+cm, data=Dataset002a)

> .Table

, , cm = コマーシャルをよく見る

	Goodsbuy(商品購入)	
gender	ある	ない
男	52	288
女	49	278

, , cm =コマーシャルを余り見ない

	goodsbuy(商品購入)	
gender	ある	ない
男	0	85
女	12	54

,, cm = コマーシャルは他に変える

	goodsbuy(商品購入)	
gender	ある	ない
男	10	94
女	8	77

これらからコマーシャルをよく見る人は商品を購入しやすいことが多重分割表から分かる。

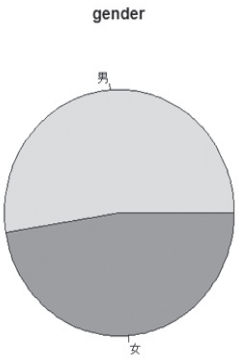
以下はRの円グラフスクリプトである。

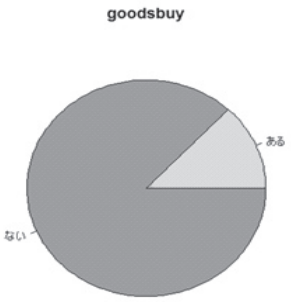
スクリプトウィンドウ

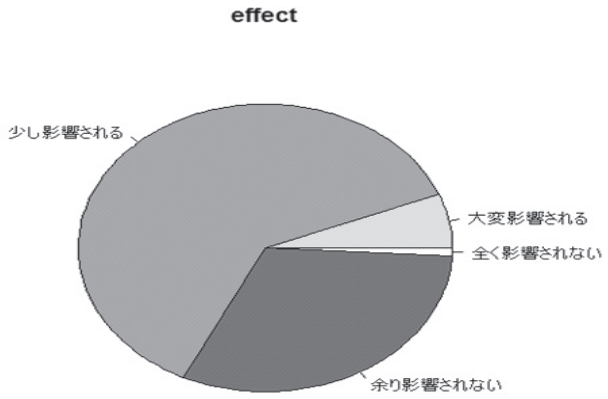
```
pie(table(Dataset0a1$gender), labels=levels(Dataset0a1$gender),
   main="gender", col=rainbow_hcl(length(levels(Dataset0a1$gender))))
pie(table(Dataset0a1$cm), labels=levels(Dataset0a1$cm), main="cm",
   col=rainbow_hcl(length(levels(Dataset0a1$cm))))
pie(table(Dataset0a1$effect), labels=levels(Dataset0a1$effect),
   main="effect", col=rainbow_hcl(length(levels(Dataset0a1$effect))))
pie(table(Dataset0a1$goodsbuy), labels=levels(Dataset0a1$goodsbuy),
  main="goodsbuy", col=rainbow_hcl(length(levels(Dataset0a1$goodsbuy))))
```

出力ウィンドウ

```
> library(colorspace, pos=4)
> pie(table(Dataset0a1$goodsbuy), labels=levels(Dataset0a1$goodsbuy),
+     main="goodsbuy", col=rainbow_hcl(length(levels(Dataset0a1$goodsbuy))))
> pie(table(Dataset0a1$gender), labels=levels(Dataset0a1$gender),
+     main="gender", col=rainbow_hcl(length(levels(Dataset0a1$gender))))
> pie(table(Dataset0a1$cm), labels=levels(Dataset0a1$cm), main="cm",
+     col=rainbow_hcl(length(levels(Dataset0a1$cm))))
> pie(table(Dataset0a1$effect), labels=levels(Dataset0a1$effect),
+     main="effect", col=rainbow_hcl(length(levels(Dataset0a1$effect))))
> pie(table(Dataset0a1$goodsbuy), labels=levels(Dataset0a1$goodsbuy),
+     main="goodsbuy", col=rainbow_hcl(length(levels(Dataset0a1$goodsbuy))))
```







性別

		度数	パーセント	有効パーセント	累積パーセント
有効	男	529	52.5	52.5	52.5
	女	478	47.5	47.5	100.0
	合計	1007	100.0	100.0	

性別 と コマーシャルを見る のクロス表

度数

		コマーシャルを見る			
		よく見る	余り見ない	コマーシャルは他に変える	合計
性別	男	340	85	104	529
	女	327	66	85	478
合計		667	151	189	1007

性別 と 品物を購入する のクロス表

度数

		品物を購入する		
		ある	ない	合計
性別	男	62	467	529
	女	69	409	478
合計		131	876	1007

§３　ロジスティク回帰分析

アンケート調査や市場調査において目的変数として商品に満足しているか、またはいないか、臨床試験において薬剤が有効であるか、ないか等、２項反応や多項反応を求めることが多い。統計的には独立性の検定でχ２検定である。

今　変数ｙを２項反応変数とする。π(y)= Pr(Y= 1)= 1 －Pr{(Y=0) | gender=I ,cm=k effect=j } とする。Y =1 for 商品購入。　Y=0 for 商品購入なし。

ロジスティク回帰モデルを

$$\pi(y)=\frac{\exp(\alpha+\beta gender+\gamma cm+\delta age+\varepsilon effect)}{1+\exp(\alpha+\beta gender+\gamma cm+\delta age+\varepsilon effect)}$$

とする。次は odds の対数を取った logit 関数で、link 関数は線形モデルに変換されている。

$$\log it(\pi(y))=\log\left[\frac{\pi(y=1)}{1-\pi(y=1)}\right]=\alpha+\beta\ \text{gender}+\gamma\ \text{cm}+\delta\ \text{age}+\varepsilon\ \text{effect}+\text{error}$$

ロジスティク回帰モデルを分析するために次のＲのスクリプトを入力する。

```
GLM.2 <- glm(goodsbuy ~ cm +effect +gender +age, family=binomial(logit),
data=Dataset0a1)
summary(GLM.2)
```

以下は出力結果である。

```
Call:
glm(formula = goodsbuy ~ cm + effect + gender + age, family = binomial(logit),
    data = Dataset0a1)
Deviance Residuals:  ＃残差
    Min       1Q   Median       3Q      Max
-2.3866   0.3321   0.4117   0.5458   1.4549
Coefficients:  回帰係数の推定値と 赤池の AIC　情報量基準である。
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.039081   0.382098  -2.719  0.00654 **
cm[T.余り見ない]                 -0.017983   0.345159  -0.052  0.95845
cm[T.コマーシャルは他に変える]    0.018195   0.298464   0.061  0.95139
effect[T.少し影響される]          2.376976   0.296886   8.006 1.18e-15 ***
effect[T.余り影響されない]        3.189443   0.385476   8.274  < 2e-16 ***
effect[T.全く影響されない]       16.186419 485.084750   0.033  0.97338
gender[T.女]                    -0.059317   0.205084  -0.289  0.77240
age                              0.024534   0.009157   2.679  0.00738 **
Signif. codes:                   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
```

Null deviance:　778.52　on 1006　degrees of freedom
Residual deviance:　673.75　on　999　degrees of freedom
AIC:　689.75
Number of Fisher Scoring iterations: 14
このRの結果から回帰係数のパラメターの推定値について仮説検定すれば、pr(>|z|)から判定するとcm とgender について帰無仮説　H_0; $\gamma=0$ と H_0; $\beta=0$　を採択する。影響について、全く影響されない　について　H_0; $\varepsilon=0$ を採択する。その他については帰無仮説を棄却して対立仮説を採択する。
結論；
アンケート調査や市場調査はCategorical Data Analysis を行うための手順でもあり、応用範囲は幅が広い。Rを活用することによってより深い分析が可能になる。標本数n が十分大きいとき、ODBC のデータベースを活用することが望ましい。

参考文献

[1] 青木繁伸、R による統計解析、Ohmsha.
[2] 荒木孝治、R と R コマンダーではじめる多変量解析、日科技連.
[3] Daniel B. Wright and Kamala London, *Modern Regression Techniques Using R*, SARGE, 2011.
[4] 船尾暢男、R コマンダーハンドブック、Ohmsha.
[5] Joseph M. Hibe, *Logistic Regression Models*. Chapmans & Hall, 2009.
[6] John Verzani, *Using R for Introductory Statistics.* Chapmans & Hall/CRC, 2008.
[7] 金明哲、R によるデータサイエンス、森北出版.
[8] 間瀬茂、R プログラミングマニュアル、数理工学社.
[9] Murray Aitkin, Brian Francis, John Hide and Ross Danrnell, *Statistical Modelling in R,* Oxford University Press, 2009.
[10] Peter Dalgaard, *Introductory Statistics with R,* Springer, 2002.
[11] Robert A. Muenchen, *R for SAS and SPSS Users*, Springer, 2009.
[12] Simon J. Sheather, *A Modern Approach to Regression with R*, Springer, 2009.