RExcel：ExcelでRを使う（2）

㈱農業・食品産業技術総合研究機構
農村工学研究所農村計画部主任研究員

合崎 英男（Aizaki Hideo）

■2000年3月北海道大学大学院農学研究科博士後期課程修了。博士（農学）。農林水産省農業研究センター研究員、農業工学研究所研究員、同主任研究官を経て、06年4月より現職。専門分野は農業経済学（主に環境配慮や食品安全性に関する意思決定分析）。

1. はじめに

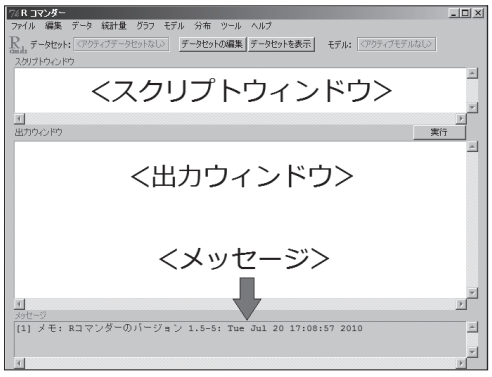
RExcelシリーズ第2回では、Rコマンダーを使った統計処理の手順について簡単に紹介します。紙幅が限られていますので、詳細な情報を知りたい方は、必要に応じて既存の文献（例えば[1][2][3][4]）も参照してください。

2. Rコマンダーの概要

RExcelを介してExcelとR、Rコマンダーを連携させるときも、Rコマンダーの画面が表示されます（図1）。Rコマンダーのメニューを Excelのツールバーと統合表示するか、Rコマンダー本体に表示するかによって、見た目が少し異なりますが（前回の図3と図4を参照）、それ以外に大きな違いはありません。Rコマンダーの画面には、3つのウィンドウがあります。メニューから実行した各種機能に対応したスクリプト（コマンド）が表示される「スクリプトウィンドウ」、実行結果が表示される「出力ウィンドウ」、各種メッセージが表示される「メッセージ」です。「メッセージ」には、分析に問題があるときの警告なども表示されますので、注意してください。

Rコマンダーが実装している機能は、「ヘルプ」を含めて9つに大別されます（図1のツールバーを参照）。「ファイル」にはRコマンダーの各ウィンドウに示される情報の保存などの機能、「編集」には各ウィンドウ上でのテキスト編集などの機能が含まれます。「データ」は、データセットの保存・読み込みや変数の操作といった機能が該当します。「統計量」にはデータセット内の

図1　Rコマンダーの画面構成



変数の要約統計量を求めたり、各種検定や統計分析を行ったりする機能が含まれます。「グラフ」には作図関連の機能、「モデル」には統計分析の適用結果に対する各種処理の機能、「分布」にはさまざまな統計分布の確率を求めたりする機能、「ツール」にはパッケージやRコマンダーのプラグインをロードする機能などが含まれます。これらの機能はバージョンの変更にともなって追加されたり、Rコマンダーのプラグインによって拡張したりすることができます。

3. 仮想例

Rコマンダーでの操作例を紹介するにあたって、1つ仮想例を設定します。ある町の住民に、環境保全に関するボランティア活動への協力意向などを問う質問紙調査を実施したと仮定します（図2）。この調査の目的は、ボランティア活動への協力意向（問1への回答）が、普段の環境配慮行動の程度(問2で選択された項目数)と性別（問3）とどのような関係にあるかを調べることとします。

図3は、30名分の仮想の回答結果をExcelで整理したものです。各列に回答者の識別番号（ID）をはじめとする変数を配置し、1行目に変数名、2行目以降の各行に回答者ごとの回答結果を入力しています。すべての設問への回答結果は、次のルールにしたがって数値として入力しています。問1（q1）では「協力したい」が「1」、「協力したくない」が「0」とします。問2（q2a ～ q2e）では、環境に配慮した行動として取り組んでいると選択した項目を「1」、選択しなかった項目を「0」として、5つの項目それぞれ別々の変数（q2a ～ q2e）としています。問3（q3）では「女性」が「1」、「男性」が「0」とします。

図2　仮想の設問

問1　あなたの住む町の環境保全のために年3回のボランティア活動への協力が求められているとします。あなたのお考えに近い選択肢を1つ選んでください。

A. 協力したい　　B. 協力したくない

問2　あなたが普段行っている環境に配慮した行動として、該当するものをすべて選んでください。

A. 長く使える商品を選ぶ。
B. エネルギー消費の少ない商品を選ぶ。
C. リサイクルされた商品を選ぶ。
D. 節水する。
E. 使っていない家電製品のコンセントは抜く。

問3　あなたの性別を教えてください。

A. 男性　　B. 女性

図3　仮想の回答結果

	A	B	C	D	E	F	G	H
1	ID	q1	q2a	q2b	q2c	q2d	q2e	q3
2	1	0	0	0	0	0	0	1
3	2	0	0	0	0	0	0	0
4	3	1	1	0	1	1	0	1
5	4	0	1	0	0	0	1	1
6	5	1	1	1	1	1	0	1
7	6	0	1	0	0	0	0	0
8	7	1	1	0	0	1	1	1
9	8	0	0	0	0	0	0	0
10	9	1	1	0	0	1	0	1
11	10	0	1	0	0	0	1	1
12	11	1	1	0	0	0	0	0
13	12	0	0	0	0	0	0	0
14	13	1	0	0	0	0	0	0
15	14	0	0	1	0	0	0	0
16	15	1	0	1	1	0	1	0
17	16	1	1	0	0	1	1	1
18	17	1	0	0	0	0	1	0
19	18	0	0	0	0	0	0	0
20	19	1	1	0	1	0	1	1
21	20	1	0	0	0	0	0	1
22	21	0	0	0	0	0	0	0
23	22	1	1	0	0	0	1	1
24	23	0	1	0	0	0	0	0
25	24	1	0	0	0	0	0	0
26	25	1	1	0	0	1	1	1
27	26	0	0	0	0	0	0	1
28	27	1	0	0	0	0	0	1
29	28	0	1	0	0	0	1	0
30	29	1	0	0	1	0	0	1
31	30	0	0	1	0	0	0	1

4. Rコマンダーでの操作例

(1)　ExcelからRへのデータセット転送

Excel上で30名分の回答結果を整理したところで、前回紹介した方法でExcelからRへデータセットを移します（前回の4.(1)を参照）。その際、データセット（データフレーム）名を「EnvSurv」とします。

⑵　変数の形式変換と新しい変数の作成

すべての設問の回答結果は、Excel 上では数値としたため（図 3）、R へ移されたデータセットに含まれる変数も数値変数とみなされます。しかし、R で統計処理する際には、カテゴリカルな回答結果については、R でいうところの「因子（factor）」に変換するのがよいでしょう。因子は整数値ベクトルで、各整数値に対応する水準名（文字）も持っています。ここでは、ボランティア活動への協力意向（q1）と性別（q3）を因子に変換します。

R コマンダーで数値変数を因子に変換するには、メニューから「データ（Data）」→「アクティブデータセット内の変数の管理（Manage variables in active data set）」→「数値変数を因子に変換（Convert numerical variables to factors）」を選択します（カッコ内は英語表示でのメニュー項目名：以下、同様）。図 4 に示す別画面が開きますので、最初に q1 を数値変数から因子に変換させます。「変数（1 つ以上選択）」で「q1」をクリックして選択し、「因子水準」を「水準名を指定」とします。そして、「新しい変数名または複数の変数に対する接頭文字列：」に「PART」と入力します。ここで［OK］ボタンを押すと、もとの数値に対応させる水準名を入力する別画面が現れます（図 5）。上述の回答選択肢と数値との対応に基づいて、「0」に「NO」、「1」に「YES」を入力して［OK］ボタンを押せば、因子への変換作業は終了です。同様に、q3 も数値変数から因子に変換させます。画面は省略しますが、新しい変数名は「GENDER」、数値と水準名との対応関係は「0」が「MALE」、「1」が「FEMALE」とします。

次に、問 2 の回答結果に該当する 5 つの数値変数（q2a ～ q2e）を使って、普段から取り組んでいる環境に配慮した行動数という新しい数値変数（ENVACT）を作成します。

図 4　数値変数の因子への変換

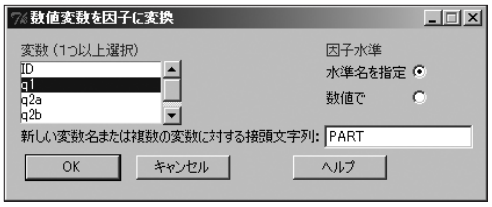
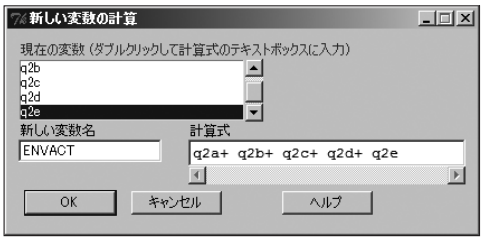


図 5　数値と水準名の対応設定

図 6　新しい変数の計算



新しい変数の作成には、R コマンダーのメニューの「データ（Data）」→「アクティブデータセット内の変数の管理（Manage variables in active data set）」→「新しい変数の計算（Compute new variable）」を選択します。別画面（図 6）が現れますので、「新しい変数名」に「ENVACT」と入力します。「計算式」に新しい変数を得るための式を入力します。今回は、q2a ～ q2e の 5 変数の合計を求めますので、「計算式」に「q2a + q2b + q2c + q2d + q2e」と入力して［OK］ボタンを押します。変数名はキーボードからも入

力できますが、画面上の「現在の変数（ダブルクリックして計算式のテキストボックスに入力）」に表示された変数名をダブルクリックしても入力できます。

以上の作業によって因子 PART と GENDER、ならびに新しい数値変数 ENVACT が、データセット EnvSurv に追加されます。画面は省略しますが、データセットに変数が追加されるたびに、R コマンダーの「メッセージ」に「メモ」として追加後のデータセットの行数・列数が表示されます。

⑶ 統計量の計算と棒グラフの作成

データセットの準備が終わりましたので、回答結果の全体像を把握してみます。

因子については、各水準の頻度を求めます。R コマンダーのメニューから「統計量（Statistics）」→「要約（Summaries）」→「頻度分布（Frequency distributions）」を選択すると、別画面（図 7）が現れます。今回は 2 つの因子の頻度を調べますので、Ctrl キーを押しながら、2 つの変数名をクリックして［OK］ボタンを押すと、出力ウィンドウに両変数の各水準の頻度と構成比率（%）が表示されます（出力は省略）。

数値変数については、平均や標準偏差などの基本的な統計量を算出します。R コマンダーのメニューから「統計量（Statistics）」→「要約（Summaries）」→「数値による要約（Numerical summaries）」を選択すると、別画面（図 8）が現れます。「変数（1 つ以上選択）」から変数 ENVACT を選択して［OK］ボタンを押せば、設定に応じて平均等が出力ウィンドウに表示されます（出力は省略）。

他方、グラフにより変数の特徴を把握するこ

図 7　頻度の計算

図 8　数値による要約

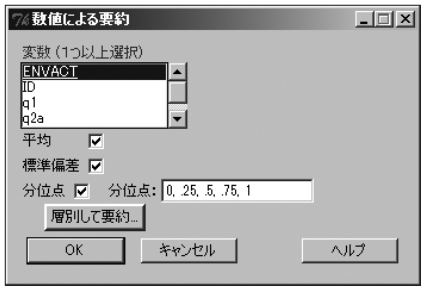


図 9　棒グラフの作成

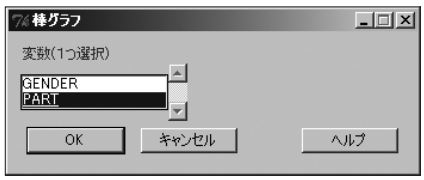
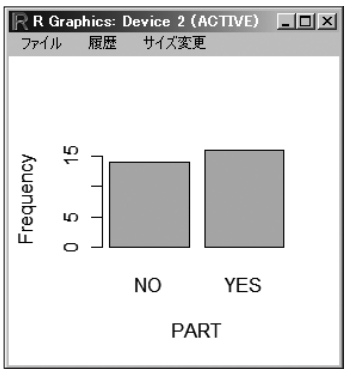


図 10　棒グラフの出力結果



ともできます。一例として、問1の回答結果(変数 PART）の棒グラフを作成します。R コマンダーのメニューから「グラフ（Graphs)」→「棒グラフ(Bar graph)」を選びます。別画面(図 9）が現れますので、「変数（1 つ選択)」で作図に使う変数 PART を選択して［OK］ボタンを押すと、別画面に棒グラフが現れます（図 10)。作成したグラフは、グラフ画面のメニューから「ファイル」→「別名で保存」として適当なファイル形式を選択すれば、画像ファイルとして保存できます。同様に「ファイル」→「クリップボードにコピー」を選択すれば、ビットマップかメタファイルとしてコピーできます。

⑷　クロス表と独立性の検定

環境保全のためのボランティア活動への協力意向と性別はともにカテゴリカル変数ですので、クロス表で両者の関係をみてみます。

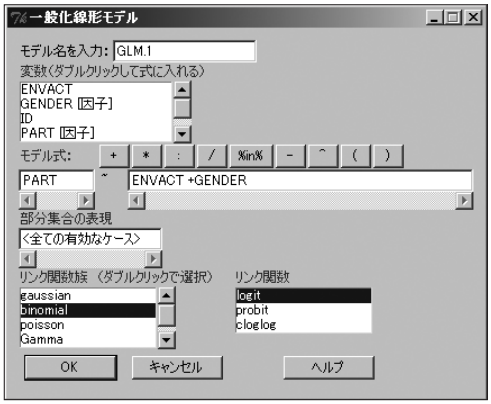
R コマンダーでクロス表（2 元表）を作成するには、メニューの「統計量 (Statistics)」→「分割表（Contingency tables)」→「2 元表（Two-way table)」を選択します。別画面（図 11）が開きますので、「行の変数（1 つ選択)」では「GENDER」、「列の変数(1 つ選択)」では「PART」を選択します。さらに、その関係を統計的に検定するときには、「仮説検定」で「独立性のカイ 2 乗検定」を選択します。これらの設定を終えて［OK］ボタンをクリックすると、出力ウィンドウにクロス表と検定結果が出力されます(出力は省略)。

⑸　2 項ロジットモデルによる分析

環境保全のためのボランティア活動への協力意向は 2 つの状態（協力したい / 協力したくない）を持ちます。このようなカテゴリカル変数

図 11　クロス表（2 元表）の作成

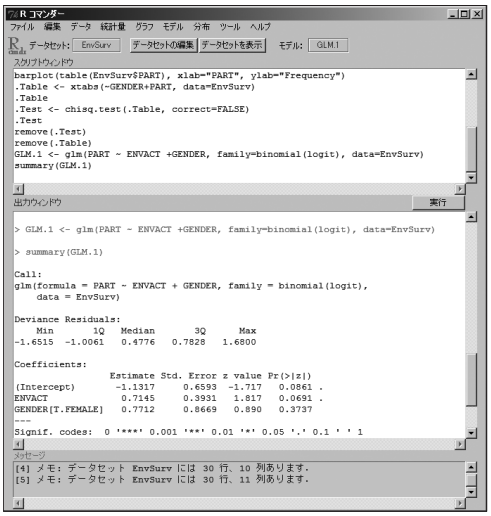
図 12　2 項ロジットモデルの設定



を分析対象とする統計手法の 1 つとして、2 項ロジットモデルがあります。この手法では、2 つの状態をとるカテゴリカル変数を目的変数とし、1 つあるいはそれ以上の説明変数を設定して、各説明変数が状態変化と関連を持つかを明らかにできます。今回は、目的変数に協力意向、説明変数に普段の環境配慮行動数と性別の 2 つを設定します。

R では、2 項ロジットモデルを一般化線型モデルの枠組みで実行します（金 [5])。R コマン

図 13　2 項ロジットモデルの出力結果



ダーのメニューから「統計量（Statistics）」→「モデルへの適合（Fit models）」→「一般化線型モデル（Generalized linear model）」を選びます。別画面（図 12）が現れますので、「モデル式」の左辺（~ の左側のボックス）に目的変数である「PART」を、同じく右辺に説明変数を「+」で結び付けながら「ENVACT + GENDER」と入力します。「リンク関数族（ダブルクリックで選択）」に「binomial」、「リンク関数」に「logit」を設定して［OK］ボタンを押すと、出力ウィンドウに分析結果が出力されます（図 13）。

ここでは 2 項ロジットモデルと表記しましたが、「ロジットモデル」「ロジスティックモデル」「ロジット回帰分析」などと表記されることもあります。

⑹　結果の保存

一通りの分析が終わったとして、実行した機能のスクリプト、出力結果、ならびに修正されたデータセットを保存します。スクリプトは R コマンダーのメニューから「ファイル（File）」→「スクリプトに名前をつけて保存（Save script as）」として保存します。出力結果も同じく「ファイル（File）」→「出力をファイルに保存（Save output as）」として保存します。データセットについては、前回の 4.⑵で紹介した手順で Excel 上に移すことができます。

なお、RExcel を利用したとしても、Excel 上と R 上の両データセットは動的に結び付いてはいません。つまり、一方のデータセットを修正しても、他方にその修正は反映されません。R 上のデータセットを修正したときには、上述のようにそれ自体、あるいはデータセットを修正したときのスクリプトを保存するのがよいでしょう。

5. おわりに

カテゴリカルデータの統計処理について詳しく知りたい方は、必要に応じて既存の文献（例えば [5][6][7][8]）を参照してください。

＊参考文献

[1] 荒木孝治編著（2007）：R と R コマンダーではじめる多変量解析：日科技連出版社 .
[2] 舟尾暢男（2008）：R Commander ハンドブック：オーム社 .
[3] 金　明哲（2006）：R コマンダー：Rcmdr(1)：ESTRELA：統計情報研究開発センター , No.150, pp.60-65.
[4] 金　明哲（2006）：R コマンダー：Rcmdr(2)：ESTRELA：統計情報研究開発センター , No.151, pp.50-55.
[5] 金 明哲（2007）：R とカテゴリカルデータのモデリング (1)：ESTRELA：統計情報研究開発センター , No.159, pp.54-59.
[6] 藤井良宜（2010）：カテゴリカルデータ解析：共立出版.
[7] 太郎丸博（2005）：人文・社会科学のためのカテゴリカルデータ解析入門：ナカニシヤ出版 .
[8] Annette J.Dobson（田中 豊・森川敏彦・山中竹春・冨田 誠訳）（2008）：一般化線形モデル入門原著第 2 版：共立出版 .