

Rによる回帰分析

奥村太一 (2015.11.16)

分析の概要

右のデータは、25人の生徒について得られた学力テスト、学習意欲尺度、学校生活適応度尺度、自尊感情尺度それぞれの得点を示したものである。¹

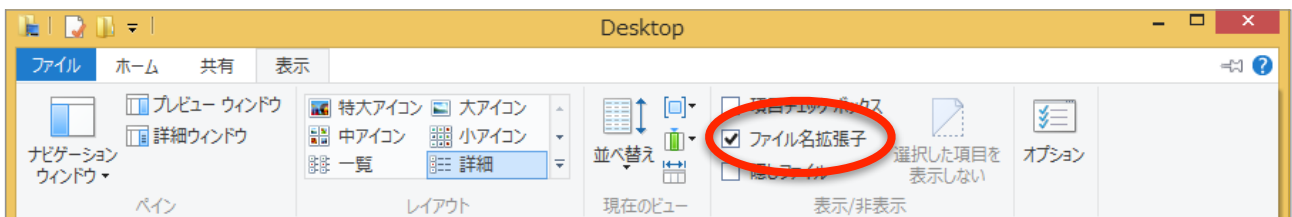
ここでは、このデータを用いて、**[学力]**を従属変数、**[意欲]**と**[適応]**を独立変数とした重回帰分析をRで実行するものとする。

	A	B	C	D	E
1	生徒	学力	意欲	適応	自尊
2	1	38	6	14	3
3	2	75	16	17	14
4	3	44	9	14	11
5	4	59	8	13	7
6	5	50	13	17	3

Rによる外部データの読み込み

Rでは、MS Excelなど他のアプリケーションで作ったデータファイルを読み込んで分析することができる。²

今、上記のデータファイルがデスクトップ上に“regpath.csv”という名で保存されているとする。³ Windowsでファイルの拡張子“.csv”が表示されない場合は、フォルダの**[表示]**タブから**[ファイル名拡張子]**にチェックを入れておく。



このデータファイルを読み込み、結果を表示するプログラムの例は、以下の通りである。ここでは、data01というオブジェクトにデータを読み込み、それを表示させている。

```
data01 <- read.csv(file="/Users/Name/Desktop/regpath.csv",header=T,
fileEncoding="Shift-JIS")
data01
```

コマンドの意味

上記プログラムに含まれる各コマンドの意味するところは以下の通り。

<-	代入を表す。A <- B で、BをAに代入せよということ。	
read.csv()	.csv形式のファイルをRに読み込ませる関数。カッコ内の引数は以下の通り。	
	file=""	ファイル名を (場所もあわせて) ""内で指定する。
	header=T	ファイルの先頭行は変数名であることを宣言する。
	fileEncoding=""	ファイルに日本語が含まれている場合、文字のエンコーディングの形式を""内で指定する。(Windows版では必要ない。)

¹ 最初の5名分だけを表示してある。なお、このデータは <http://www.juen.ac.jp/lab/okumura/data.html> から“regpath.csv”としてダウンロード可能である (リンクを右クリックして保存)。

² プログラムに直接データを打ち込むことも可能であるが、標本サイズが大きくなるとプログラムが煩雑になるのでお勧めしない。

³ CSV形式のファイルは、Excelで**[ファイル]** → **[名前をつけて保存]** → **[フォーマット]** → **[CSV (コンマ区切り) (.csv)]** で作成できる。

プログラムの説明

上記のプログラムは、1行目から2行目が、「`/Users/Name/Desktop`にある `regpath.csv` という CSV形式のファイルを読み込み、`data01`というオブジェクトを作ってそこに代入しなさい。ただし、`regpath.csv` の先頭行は変数名であり、日本語はShift-JIS形式でコーディングされています。」という内容になる。⁴

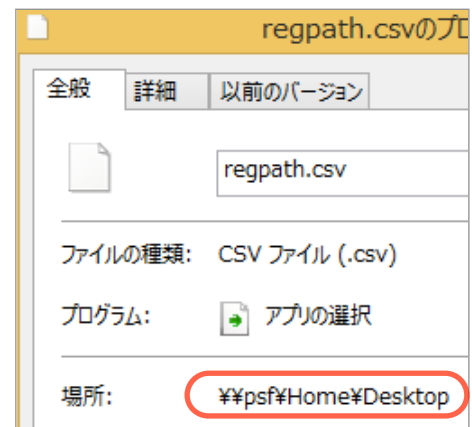
なお、3行目は、「`data01`に読み込んだデータの内容をそのまま表示しなさい。」という意味である。⁵

ファイル名の指定

Rでデータを読み込ませるとき、最もつまづきやすいのが `read.csv()` 内の `file=""` でのファイル名の指定である。この "" 内には、ファイル名だけでなく、そのファイルがPCのどの場所 (フォルダ) に保存されているのかという情報を入れなくてはいけない。

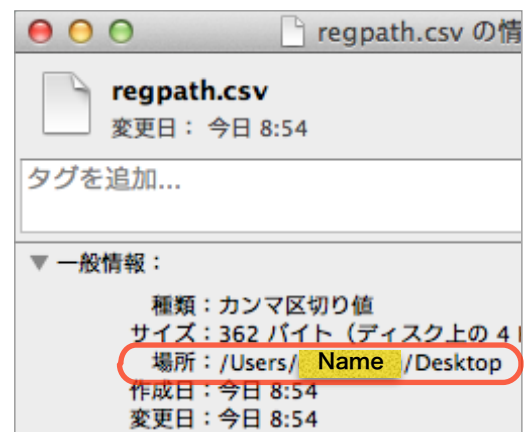
[Windows版の場合]

- ・ ファイルのアイコンを右クリック → [プロパティ] を選択。
- ・ プロパティが開いたら、[場所]にファイルの場所 (パス) が表示されるので、これを選択してコピー。
- ・ これを `file=""` の中に貼り付け。ただし、**¥は/に置き換える**こと。最後に、**/ファイル名.csv** を追加。
- ・ 今の場合、`file="//psf/Home/Desktop/regpath.csv"` となる。



[Mac版の場合]

- ・ ファイルのアイコンを2本指クリック → [情報を見る] を選択。
- ・ [場所]にファイルの場所 (パス) が表示されるので、これを選択してコピー。
- ・ これを `file=""` の中にそのまま貼り付け。最後に、**/ファイル名.csv** を追加。
- ・ 今の場合、`file="/Users/Name/Desktop/regpath.csv"` となる。⁶



以上のプログラムを実行すると、以下のように、読み込んだデータの内容が表示される。もとの CSVファイルの内容と比較して、正常に読み込まれているか確認しよう。

⁴ 本来は1~2行目全体で1つのコマンドであるが、長いので途中で改行している。また、ここでは代入するオブジェクト名を `data01` としているが、任意でよい。

⁵ この部分は、`print(data01)`としてもよい。また、先頭の数行だけ表示させたい場合は、`header(data01)` とすればよい。

⁶ 一部画像を加工しています。

```

> data01 <- read.csv(file="/Users/Name/Desktop/regpath.csv",header=T,
+                   fileEncoding="Shift-JIS")
> data01
  生徒 学力 意欲 適応 自尊
1     1   38    6   14    3
2     2   75   16   17   14
3     3   44    9   14   11
4     4   59    8   13    7
5     5   50   13   17    3
(…以下省略)

```

回帰分析の実行

以上のプログラムを実行すれば、data01には変数名とともにそのデータが読み込まれている。このデータを用いて、[学力]を従属変数、[意欲]と[適応]を独立変数とする重回帰分析を実行し、その結果を表示するプログラムは以下の通りである。

```

fit01 <- lm(学力~意欲+適応, data=data01)
summary(fit01)

```

コマンドの意味

上記プログラムに含まれる各コマンドの意味するところは以下の通り

lm()	重回帰分析を実行する関数。カッコ内の引数の意味は以下の通り。	
	y~x1+x2+…	yが従属変数、x1+x2+…が独立変数であることを表す。
	data=	分析の対象となるデータを = 以下で指定する。
summary()	分析結果の要約を表示するための関数。	

プログラムの説明

上記のプログラムは、1行目が「学力を従属変数、意欲と適応を独立変数とした重回帰分析を行いなさい、ただしデータは先ほど作成したdata01を用いなさい。分析結果は、fit01というオブジェクトを作ってそこに代入しなさい。」という内容になる。⁷

2行目は、「fit01に代入された重回帰分析の結果を要約して表示しなさい。」という意味である。

以上のプログラムを実行すると、以下のように分析結果が表示されるはずである。⁸

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.0217     12.7717   0.628  0.5364
意欲         2.4351      0.9204   2.646  0.0148 *
適応         1.6665      1.0069   1.655  0.1121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.02 on 22 degrees of freedom
Multiple R-squared:  0.4192, Adjusted R-squared:  0.3663
F-statistic: 7.938 on 2 and 22 DF, p-value: 0.002539

```

⁷ ここでは、代入するオブジェクト名をfit01としているが、任意でよい。また、もしデータの読み込みの際に、data01と異なるオブジェクト名で代入していたとしたら、data=data01の部分はそれに合わせて変更する必要がある。

⁸ アウトプットのうち、最初の数行は省略してある。

分析結果 (切片と偏回帰係数)

切片および偏回帰係数⁹の推定値と検定結果は、Coefficients: 以降の表に提示されている。

行については、(Intercept) が切片、それ以降は各独立変数にかかる偏回帰係数に関する部分である。

列には、以下の表のように切片と偏回帰係数の統計的推測の結果に関する情報が並んでおり、有意確率の値に応じて右側に '*' などの記号が振られている。

Estimate	推定値 (データから計算された切片と偏回帰係数の値)
Std. Error	推定の標準誤差 (区間推定を行うときに利用する)
t value	切片と偏回帰係数それぞれの検定における検定統計量 t の値
Pr(> t)	切片と偏回帰係数それぞれの検定における有意確率 (p 値)

今の例では、回帰式の切片が8.02、[意欲]の偏回帰係数が2.44、[適応]の偏回帰係数が1.67となっている。すなわち、

- ・ **切片**：[意欲]と[適応]がともに0点の生徒については、[学力]の予測値は8.02点
- ・ **[意欲]の偏回帰係数**：[適応]の得点と同じで[意欲]が1点違う生徒同士では、[学力]の予測値が2.44点違う (前者が高い生徒ほど後者も高い)
- ・ **[適応]の偏回帰係数**：[意欲]の得点と同じで[適応]が1点違う生徒同士では、[学力]の予測値が1.67点違う (前者が高い生徒ほど後者も高い)

と解釈できる。

ただし、検定の結果5%水準で有意なのは[意欲]の偏回帰係数のみである。切片および[適応]の偏回帰係数については、これらの値が母集団では0であるという帰無仮説はいずれも棄却されない。なお、これらの検定は、検定統計量 t とその自由度 (=人数-独立変数の数-1) を用いて行われる。結果の下から3行目 (22 degrees of freedom) にあるとおり、この場合の自由度は22である。従って、[意欲]の偏回帰係数については、例えば

「“意欲”の偏回帰係数は5%水準で統計的に有意であった ($b = 2.44, t(22) = 2.65, p = .015$)。 」

といった形で検定結果を記載することになる。¹⁰

分析結果 (決定係数)

結果の下から2行目に決定係数 (Multiple R-squared:) と自由度調整済み決定係数 (Adjusted R-squared:) の値が報告されている。今の場合、それぞれの値はおよび $R^2 = .42$ および $R^2_{adj} = .37$ となっている。

なお、結果の最下段には決定係数の検定結果として、検定統計量 F と2つの自由度、有意確率の値が記載されている。この例では、検定結果は5%水準で統計的に有意となっており、母集団において決定係数がゼロであるという帰無仮説は棄却されることになる ($F(2, 22) = 7.94, p = .003$)。

⁹ 単回帰分析の場合は回帰係数と呼ぶのが一般的であるが、双方を記載すると煩雑なためここでは偏回帰係数に統一する。

¹⁰ 偏回帰係数は記号 b で表すことが多い。