人間行動システム統計演習 A

第 4 回：相関と回帰　担当：中川、寺井、和嶋 *

2015 年 5 月 7 日

1　相関

相関係数とは、二変数間の関係を数値で示す統計的手法です。-1 から 1 の間の実数値をとり、1 に近いとき 2 つの変数には正の相関があるといい、-1 に近いとき負の相関があるといいます。0 に近いときは相関が弱いといいます。1 もしくは-1 となる場合、2 つの変数は線形従属の関係にあります。（まず始めに、`shintai.csv` を `shintai` という名前のデータセットとして読みこんで下さい。復習として、身長を X 軸と体重を Y 軸として、「最小 2 乗直線」のオプションのみをつけて、それらの散布図を描いてください（「グラフ」→「散布図」）。）

1.1　相関行列を求め方

二変数間の相関係数を求めるためには、「統計量」→「要約」→「相関行列」をクリックし、対象となる変数を選択して OK をクリックします。「相関のタイプ」オプションとして「ピアソンの積率相関係数」、「スピアマンの順位相関」、「偏相関」のいずれかをクリックしてください。一般的に「相関係数」と言った場合、「ピアソンの積率相関係数」を意味します。「ノンパラメトリック密度推定」オプションは、欠損値がある場合の計算方法に関するオプションです。「ペアワイズの p 値」オプションにチェックを入れると、相関係数の有意性検定における p 値を出力します。結果として、選択した変数に関する相関係数を要素として持つ「相関行列」が出力されます。対角要素は同一の変数間の相関係数を表すため `1.0000` となっています。R コマンダーにて女子のデータのみ、またはクラス 1 の人のみを対象として分析を行いたい場合は、事前に、対象となるデータからなるデータセットを作成し、分析を実施してください（「データ」→「アクティブデータセット」→「アクティブデータセットの部分集合を抽出」）。

* 御質問などは寺井 (terai.a.aa@m.titech.ac.jp) にメールにてご連絡下さい。

1.2 相関係数の検定

相関行列を求める際に「ペアワイズの p 値」オプションにチェックを入れる相関係数の有意性検定における p 値を出力しますが、「統計量」→「要約」→「相関の検定」により、cor.test 関数を用いて相関係数の有意性検定と区間推定も同時に行い、その結果を表示してくれます。「母相関係数が 0 である」という帰無仮説に関して検定を行うことを、「相関係数の有意性検定」といいます。このとき、

帰無仮説：「母相関係数は 0 と等しい（$\rho = 0$）」

対立仮説：「母相関係数は 0 と等しくない（$\rho \neq 0$）」

となります。こちらで計算することができる相関係数は、ピアソンの積率相関係数、スピアマンの順位相関係数、ケンドールのタウ（順位相関係数）です。一般的に相関係数の検定では「両側」検定が用いられます。

身長と体重の相関係数に関し、検定を行ってみましょう。

```
  Pearson's product-moment correlation

data:  shintai$height and shintai$weight
t = 8.0984, df = 78, p-value = 6.067e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.5356690 0.7797683
sample estimates:
      cor
0.6758432
```

t = 8.0984 として出力されている値は、t 統計量です。df = 78 は自由度を表します（相関係数を求める際の自由度は df = n-2 です)。p-value = 6.067e-12 は p 値を表します。この場合、6.067e-12<0.01 のため 1%(0.01) より小さいため、「母相関係数が 0 である」という帰無仮説は棄却され「母相関係数は 0 でない」という対立仮説が採択されます [*1]。すなわち、「得られた相関係数 (cor の下に書かれている値）0.6758432 は有意である」といえます。95 percent confidence interval の下に書かれている数値は、母相関係数の 95% 信頼区間を表します。

オプションとして選択できる各相関係数は以下のような意味を持ちます。

[*1] この値が 0.05 以下の場合は 5% 水準で有意な差が、0.01 以下の場合は 1% 水準で有意な差があることを意味します。

- ピアソンの積率相関係数
  一般的に、「相関係数」と言った場合、この相関係数を意味します。
- スピアマンの順位相関係数・ケンドールのタウ
  順位データから求められる相関の指標。
- 偏相関係数
  対象となる２変数以外変数の影響を取り除いた相関係数（例：ポストの数、事故数の関係を検討する上で、人口の影響を取り除く）。

1.3 cor.test 関数の出力制御

cor.test 関数は多くの結果を返してくれます。相関係数のみを求めたい場合は、cor 関数を利用するほうが簡単ですが、cor.test 関数の場合、以下の手順で表示する結果を細かく指定することができます。スクリプトウィンドウにて実行してみて下さい。

```
> corRes <- cor.test(shintai$height, shintai$weight)↵
> str(corRes)↵
List of 9
 $ statistic  : Named num 8.1
  ..- attr(*, "names")= chr "t"
 $ parameter  : Named num 78
  ..- attr(*, "names")= chr "df"
 $ p.value    : num 6.07e-12
 $ estimate   : Named num 0.676
  ..- attr(*, "names")= chr "cor"
 $ null.value : Named num 0
  ..- attr(*, "names")= chr "correlation"
 $ alternative: chr "two.sided"
 $ method     : chr "Pearson's product-moment correlation"
 $ data.name  : chr "shintai$height and shintai$weight"
 $ conf.int   : atomic [1:2] 0.536 0.780
  ..- attr(*, "conf.level")= num 0.95
 - attr(*, "class")= chr "htest"

> corRes$estimate↵ # 相関係数のみの表示
      cor
0.6758423
```

上記では、最初に cor.test の結果に corRes というオブジェクト名をつけて格納しています。オブジェクトの構造を表示するには、str( ) を利用します。この結果から、corRes$statistic, corRes$parameter, corRes$p.value, corRes$estimate などとすることにより、それぞれ t 統計量、自由度、p 値、相関係数にアクセスできることがわかります。cor.test 関数はこれらの情報を整形し、まとめて表示してくれるわけです。

1.4 相関算出の際の注意事項

相関を算出する際は、以下の点に留意する必要があります。

- 相関係数は二変数の関係を計測しているに過ぎないため、変数間の因果関係を説明するものではない。
- 相関係数は順序尺度であり間隔尺度ではないので、例えば「相関係数が 0.2 と 0.4 であることから、後者は前者より 2 倍の相関がある」などと言うことはできない。
- いくつかのはずれ値が存在する場合には，それらの値に引きずられて、不当に高い相関係数が得られる。このような場合には順位相関係数を使用したほうがよい。
- データが複数個のグループから構成されている場合には、全体をまとめた相関係数の値は個々のグループにおける相関係数とは異なったものになることがある。
- 曲線相関の場合には順位相関係数を使用したほうがよい。極端な場合であるが $Y = X^2$ の関係がある場合に、順位相関係数は 1 になるが、ピアソンの積率相関係数は 1 にはならない。
- スピアマンの順位相関係数とケンドールの順位相関係数は、定義に差があるだけで、どちらも 2 変数の順位が一致するか否かの指標として用いられる。

演習

女子のデータ (sex="F") のみ、男子のデータのみ (sex="M") に関する身長 (height) と体重 (weight) の相関係数をそれぞれ求めて下さい。

2 母相関係数が 0 以外の特定の値であるかどうかの検定

ピアソンの積率相関係数に関し、「母相関係数が 0 以外の値特定の値である」という帰無仮説に関して検定を行う方法を紹介します。この検定を行うための関数は、標準の R には用意されていません。従って、R でこのような検定を行うには、検定の定義式に基づいて自作関数を作成するか、インターネット等で公開されている関数を使用する必要があります。

標本数（データの組数）n、（標本）相関係数を r とします。このとき、「母相関係数 ρ が 0 以外の特定の値 $x(\neq 0)$ である」という帰無仮説に関する検定を行います。この場合,
帰無仮説 H_0：「母相関係数は x と等しい（$\rho = x(\neq 0)$）」
対立仮説 H_1：「母相関係数は x と等しくない（$\rho \neq x(\neq 0)$）」

となります。上記の帰無仮説のもとで、標本相関係数 r と母相関係数 ρ の差の標本分布が分散 1、平均 0 の標準正規分布に従うように変換します。
まず、標本相関係数、母相関係数の z 変換値 ξ_r,ξ_ρ を求めます。

$$\xi_r = \frac{1}{2} ln(\frac{1+r}{1-r})$$
$$\xi_\rho = \frac{1}{2} ln(\frac{1+\rho}{1-\rho})$$

次に、検定統計量 z_0 を求めます。z_0 は、標準正規分布に従う値です。

$$|z_0| = \frac{|\xi_r - \xi_\rho|}{\sqrt{\frac{1}{n-3}}}$$

この z_0 の値の確率を計算し、その確率が十分に低ければ（1% または 5% より低ければ)、帰無仮説を棄却し「母相関係数は $x(\neq 0)$ と有意な差がある」と結論づけることができます [*2]。以下の手順で、身長と体重の母相関係数が 0.5 と有意な差があるか否かを検定します。スクリプトウィンドウにコマンドを記載し、実行してみてください。

```
 > zr <- atanh(corRes$estimate) ↵
> zrho <- atanh(0.5) ↵
> z <- abs(zr-zrho)/sqrt(1/( corRes$parameter+2 - 3)) ↵
> p <- 2*pnorm(z, mean=0, sd=1, lower.tail=FALSE)↵
> z  ↵
 [1]  2.38779386
> p  ↵
 [1]  0.01694985
```

atanh(x) は、$\frac{1}{2} ln \frac{1+x}{1-x}$ を返す関数、abs() は絶対値を返す関数、pnorm() は、正規分布における特定の値の確率を返す関数です。ここでは、標準正規分布なので平均 0(mean=0)、標準偏差 1(sd=1) を指定しています。

corRes$estimate は相関係数を表しています。自由度は標本数-2 となっているため、ここでは corRes$parameter+2 とすることで標本数を求めています。z = 2.38779386 として z 統計量、p = 0.01694985 として p 値が出力されています。p が 0.05 以下であることから身長と体重の母相関係数と 0.5 に 5% 水準で有意な差があるという結果が得られています。

[*2] 詳細は以下の Web ページを参照してください
http://aoki2.si.gunma-u.ac.jp/lecture/Corr/corr2.html

3 回帰分析

2 変数 X（独立・説明変数）と Y（従属・目的変数）の間の関係を $Y = \alpha + \beta X$ として、α(切片)、β（回帰係数）の推定を行い、目的変数 Y が説明変数 X によってどれくら説明できるかを定量的に分析します。[*3] 説明変数が 1 つの場合を「単回帰分析」、複数個の場合を「重回帰分析」とよびます。「統計量」→「モデルへの適合」→「線形回帰」をクリックし、目的変数、説明変数を選択してください。女子のデータのみ、またはクラス 1 の人のみを対象として分析を行いたい場合は、「部分集合の表現」という部分に、条件式を記載してください (例：sex == “ F ”)。

身長を目的変数、体重を説明変数として分析を実施してください。出力ウィンドウに以下のような結果が出力されます。

```
Call:
lm(formula = height ~ weight, data = shintai)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9925  -3.8123   0.4476   3.1975  14.3058

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 135.77312    3.37018  40.287     < 2e-16 ***
weight        0.47740    0.05895   8.098    6.07e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.53 on 78 degrees of freedom
Multiple R-Squared: 0.4568,    Adjusted R-squared: 0.4498
F-statistic: 65.58 on 1 and 78 DF,  p-value: 6.067e-12
```

Residuals は残差の最小値、25% タイル点、中央値、75% タイル点、最大値を表しています。次に Coefficients として推定された、推定された切片、回帰係数に関する情報がまとめられています。Estimate として、135.77312 が切片 a の値、0.47740 が回帰係数 b の値です。結果より、身長 $= 0.47740 *$ 体重 $+ 135.77312$ であることが推定されています。また、Std.Error として切片・回帰係数の標準誤差、t value と Pr(>|t|) として「(母) 回帰係数（または切片）が 0 である」

[*3] 本授業では目的変数、説明変数がともに量的変数の場合の分析についてを扱います。

ことを帰無仮説とした検定結果が示されています。回帰係数 b に関する検定結果は、p 値が 0.01 以下であり、1% 水準で棄却されています。そのため、「(母) 回帰係数は 0 ではない」という対立仮説が採択されます。

また、最後の 3 行に回帰分析（回帰モデル）が妥当かどうかを定量的に確かめるための検定結果が示されています。Residual standard error として残差の標準誤差と自由度 (n-説明変数の数-1) が示されています。また、Multiple R-squared として決定係数（R^2 値)、Adjusted R-squared として自由度調整済決定係数が示されます。決定係数とは、説明変数が目的変数をどれくらい説明できるかを表す値です。回帰モデルのあてはまりの良さを表す尺度として利用されます。一般的な回帰分析の場合、決定係数は目的変数と推定値の相関係数を 2 乗した値になります。F-statistic として「全ての係数が 0 である（回帰モデルが妥当ではない)」という帰無仮説に関する F 検定の結果を表しています。この結果では、検定統計量 F 値=65.58、検定に用いた F 分布の自由度 (説明変数の数、n-説明変数の数-1)、p 値が示されています。

回帰分析は、`lm(`**`データセット名$目的変数名 ~ データセット名$説明変数名`**`)` または `lm(`**`目的変数名 ~ 説明変数名`**`, data=`**`データセット名`**`)` で実行できます。ただし、これで返される結果は極めて簡潔であるため、`lm( )` 関数の結果をオブジェクトに格納し、これに `summary` 関数を適応するといった手続きを経ることで詳細な結果を得ています。`RegModel.1 <- lm(height ~ weight, data=shintai)` が、回帰分析結果を `RegModel.1` というオブジェクトに格納している部分です。`summary(RegModel.1)` が、結果の表示を指示している部分です。

「モデル」→「アクティブモデルを選択」で対象となる回帰分析結果を選択し、計算結果をデータとして保存したり、残差等に関するグラフを作成することができます。

3.1 回帰分析を実施する際の注意事項

回帰分析を実施する際は、以下の点に留意する必要があります。

- 回帰分析（特に単回帰分析）では、目的変数と説明変数の「相関関係」によって回帰係数が決定されています。相関係数に関する注意事項の多くが回帰分析における注意事項となります。(因果関係を表しているわけではない、外れ値の影響、グループの影響、直線的な関係しか見る事が出来ない（ただし非線形回帰という方法もあります)）
- 説明変数を増やせば増やすほど決定係数の値は高くなりますが、回帰係数の解釈は困難になります。重回帰分析においては、モデルのあてはまりの良さを表す指標としてパラメータの個数（説明変数の個数）を考慮した指標（例 AIC：赤池情報量規準「モデル」→「赤池情報量規準 (AIC)」) があります。
- 重回帰分析の場合、説明変数間の相関が高い場合「多重共線性」という問題が生じます。この場合、回帰係数を解釈する事自体が無意味になります。

演習

生徒の身長 (height) を目的変数、両親の身長（PaHeight, MaHeight) を説明変数として重回帰分析を行ってください。

4 関数を自作する

R では、cor.test や atanh など標準で実装されている関数がたくさんありますが、自分で関数を作成する（定義する）ことができます。

2 つの値の和差積商を計算する関数

```
my_cal <- function(x,y)
{
p <- x + y
m <- x - y
t <- x * y
d <- x / y
return(c(p,m,t,d))
}
```

`my_cal <- function(x,y)` は、2 つの入力（x,y）に対して値を返す「`my_cal`」という関数を定義するという宣言文です。関数の中で行う処理は`{ }`で囲みます。`return()` で関数が返す値を指定します。ここでは、各演算の結果が入った変数をベクトルにして返り値としています。
メモ帳などで上記のプログラムを入力し、`my_cal.R` という名前で保存してください。

自作した関数を利用するときには、「ファイル」→「スクリプトファイルを開く」で読み込み、実行して下さい。一度、関数を読み込むと R を終了するまで有効です。

```
> my_cal(2,3) ↵
[1] 5.0 1.0 6.0 1.5
> my_cal(100,10) ↵
[1] 110   90 1000   10
```

`structure` を使用して、返り値にちょっとした説明をつけることも可能です。

2 つの値の和差積商を計算する関数 (返り値説明つき)

```
my_cal <- function(x,y)
{
p <- x + y
```

```
m <- x - y
t <- x * y
d <- x / y
return(structure(c(p,m,t,d),names=paste(c("+","-","*","/"))))  # 書き換え
}
```

```
> my_cal(2,3) ↵
        +         -         *         /
5.0000000 -1.0000000  6.0000000  0.6666667
```

5 本日の課題

1. 母相関係数が 0 以外の特定の値であるかどうかの検定を行う関数を作成しなさい。返り値には、どの値が z, p であるかが分かるように説明を付けること。ファイル名は `rtest.R` とする。
2. 自作した関数 `rtest.R` を使い、女子の身長と体重のピアソンの積率相関係数に関し、母相関係数が 0.5 と有意な差があるか否かを検定する (z 値、p 値を求める)
3. 男子の身長をその母親の身長で説明する回帰分析を行う。
4. 3 の回帰分析結果を利用して、従属変数と従属変数の予測値の相関の 2 乗を求める（R^2 値と同じ値になっていることを各自確認する)。

作成したスクリプトファイル・自作関数ファイル（自作関数の内容を、スクリプトファイルの先頭にコピー&ペーストし、一つのファイルにしてください）を OCWi から提出してください。課題の締め切りは次回の講義の前日です（5 月 13 日 23:59:59)。