エビデンスベーストヘルスケア特講 I (7) 相関と回帰

中澤　港（国際保健学領域・教授）

2013 年 5 月 29 日

目次

問い合わせ先：神戸大学大学院保健学研究科国際保健学領域・教授　中澤　港
e-mail: minato-nakazawa@umin.net

2012 年 6 月 23 日：第 1 版（EZR 用に書き換え）
2013 年 5 月 29 日：第 1.1.1 版（微修正）

1　２つの量的な変数間の関係

２つの量的な変数間の関係を調べるための，良く知られた方法が２つある。相関と回帰である。いずれにせよ，まず散布図を描くことは必須である。

MASS ライブラリの survey データフレームで，身長と利き手の大きさ（親指の先端と小指の先端の距離）の関係を調べるには，R コンソールでは，`require(MASS)` として MASS ライブラリをメモリに読み込んだ後であれば，`plot(Wr.Hnd ~ Height, data=survey)` とするだけである。もし男女別にプロットしたければ，`pch=as.integer(Sex)` というオプションを指定すれば良い。

EZR では，「ファイル」「パッケージに含まれるデータを読み込む」から左の枠の MASS でダブルクリックし，右の枠に現れるデータフーレムの下の方へスクロールして survey でダブルクリックしてから OK ボタンをクリックした後に，「グラフ」「散布図」と選び，x 変数として Height を，y 変数として Wr.Hnd を選び，“最小 2 乗直線” の左側のチェックボックスのチェックを外し，[OK] をクリックする。男女別にプロット記号を変えたい場合は，「層別のプロット」というボタンをクリックし，層別変数として Sex を選んで [OK] をクリックし，元のウィンドウに戻ったら再び [OK] をクリックすればよい。

1.1　相関と回帰の違い

大雑把に言えば，相関が変数間の関連の強さを表すのに対して，回帰はある変数の値のばらつきがどの程度他の変数の値のばらつきによって説明されるかを示す。回帰の際に，説明される変数を（従属変数または）目的変数，説明するための変数を（独立変数または）説明変数と呼ぶ。２つの変数間の関係を予測に使うためには，回帰を用いる。

1.2　相関分析

一般に，2 個以上の変量が「かなりの程度の規則正しさをもって，増減をともにする関係」のことを相関関係 (correlation) という。相関には正の相関 (positive correlation) と負の相関 (negative correlation) があり，一方が増えれば他方も増える場合を正の相関，一方が増えると他方は減る場合を負の相関と呼ぶ。例えば，身長と体重の関係は正の相関である。

散布図で相関関係があるように見えても，見かけの相関関係 (apparent correlation) であったり[*1]，擬似相関 (spurious correlation) であったり[*2]することがあるので，注意が必要である。

相関関係は増減をともにすればいいので，直線的な関係である必要はなく，二次式でも指数関数でもシグモイドでもよいが，通常，直線的な関係をいうことが多い（指標はピアソンの積率相関係数）。曲線的な関係の場合，直線的になるように変換したり，ノンパラメトリックな相関の指標（順位相関係数）を計算する。順位相関係数としてはスピアマンの順位相関係数が有名である。

ピアソンの積率相関係数 (Pearson's Product Moment Correlation Coefficient) は，r という記号で表し，２つの変数 X と Y の共分散を X の分散と Y の分散の積の平方根で割った値であり，範囲は $[-1, 1]$ である。最も強い負の相関があるとき $r = -1$，最も強い正の相関があるとき $r = 1$，まったく相関がないとき（2 つの変数が独立なとき），$r = 0$ となることが期待される。X の平均を $\bar{X}$，Y の平均を $\bar{Y}$ と書けば，次の式で定義される。

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

相関係数の有意性の検定においては，母相関係数がゼロ（＝相関が無い）という帰無仮説の下で，実際に得られている相関係数よりも絶対値が大きな相関係数が偶然得られる確率（これを「有意確率」という。通常，記号 p で表すので，「p 値」とも呼ばれる）の値を調べる。偶然ではありえないほど珍しいことが起こったと考えて，帰無仮説が間違っていたと判断するのは有意確率がいくつ以下のときか，という水準を有意水準といい，検定の際には予め有意水準を（例えば 5% と）決めておく必要がある。例えば $p = 0.034$ であれば，有意水準 5% で有意な相関があるという意思決定を行なうことができる。p 値は，検定統計量

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

が自由度 $n-2$ の t 分布に従うことを利用して求められる。

[*1] 例）同業の労働者集団の血圧と所得。どちらも一般に加齢に伴って増加する。

[*2] 例）ある年に日本で植えた木の幹の太さと同じ年に英国で生まれた少年の身長を 15 年分，毎年 1 回測ったデータには相関があるようにみえるが，直接的な関係はなく，どちらも時間経過に伴って大きくなるために相関があるように見えているだけである。

散布図を描いた survey データフレームの身長と利き手の大きさの間でピアソンの相関係数を計算し，その有意性を検定するには，R コンソールでは次の 1 行を打てばよい（スピアマンの順位相関について実行したい時は，methos=spearman を付ける）。

```
cor.test(survey$Height, survey$Wr.Hnd)
```

EZR では，「統計解析」の「連続変数の解析」から「相関係数の検定（**Pearson** の積率相関係数）」を選び，変数として Height と Wr.Hnd を選ぶ（**Ctrl** キーを押しながら変数名をクリックすれば複数選べる）。検定については「対立仮説」の下に「両側」「相関＜０」「相関＞０」の３つから選べるようになっているが，通常は「両側」でよい。**OK** をクリックすると，**Rcmdr** の出力ウィンドウに次の内容が表示される。

```
Pearson's product-moment correlation

data:  survey$Height and survey$Wr.Hnd
t = 10.7923, df = 206, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5063486 0.6813271
sample estimates:
      cor
0.6009909
```

これより，身長と利き手の大きさの関係について求めたピアソンの積率相関係数は，$r = 0.60$（95% 信頼区間が $[0.50, 0.69]$）であり[*3]，p-value < 2.2e-16（有意確率が 2.2×10^{-16} より小さいという意味）より，「相関が無い」可能性はほとんどゼロなので，有意な相関があるといえる。なお，相関の強さは相関係数の絶対値の大きさによって判定し，伝統的に 0.7 より大きければ「強い相関」，0.4〜0.7 で「中程度の相関」，0.2〜0.4 で「弱い相関」とみなすのが目安なので，この結果は中程度の相関を示すといえる。

男女別に相関係数の検定を実行するには，いろいろなやり方があるが，最も単純に考えれば，データセットそのものを男女別の部分集合に分け，それぞれについて分析すればよい。R コンソールでは次の 4 行を打つ（その前に，MASS ライブラリをメモリに読み込んでおかねばならないのは当然である）。

```
males <- subset(survey, Sex=="Male")
cor.test(males$Height, males$Wr.Hnd)
females <- subset(survey, Sex=="Female")
cor.test(females$Height, females$Wr.Hnd)
```

EZR では，「アクティブデータセット」の「行の操作」の「指定した条件を満たす行だけを抽出したデータセットを作成する」を選び，表示されるウィンドウで，「すべての変数を含む」はチェックが入ったまま，「サンプルを抽出する条件式」のボックスに Sex=="Male"と入力し，「新しいデータセットの名前」に Males（既にある名前と重複しなければ何でもよい）と入力して **[OK]** ボタンをクリックすると，男性だけのデータフレーム Males ができてアクティブになる。ここで先ほどと同じ「統計解析」「連続変数の解析」「相関係数の検定（**Pearson** の積率相関係数」をすれば男性の身長と利き手の大きさについてピアソンの積率相関係数を求めて有意性の検定をすることができる。
女性について同じことをするには，まず「アクティブデータセット」の下の Males と表示されている部分をクリックして出てくるウィンドウで survey を選び直し，先ほどと同じ「アクティブデータセット」の「行の操作」の「指定した条件を満たす行だけを抽出したデータセットを作成する」の「サンプルを抽出する条件式」のボックスに Sex=="Female"，「新しいデータセットの名前」で Females として **OK** ボタンをクリックしてから「統計解析」「連続変数の解析」「相関係数の検定（**Pearson** の積率相関係数」を実行すればよい。

[*3] 95% 信頼区間の桁を丸めて示す場合，真の区間を含むようにするために，四捨五入ではなく，下限は切り捨て，上限は切り上げにするのが普通である。

順位相関係数の定義

なお，スピアマンの順位相関係数 ρ は[a]，値を順位で置き換えた（同順位には平均順位を与えた）ピアソンの積率相関係数と同じである。X_i の順位を R_i，Y_i の順位を Q_i とかけば，

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$$

となる。スピアマンの順位相関係数がゼロと差がないことを帰無仮説とする両側検定は，サンプル数が 10 以上ならばピアソンの場合と同様に，

$$T = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

が自由度 $n-2$ の t 分布に従うことを利用して行うことができる。ケンドールの順位相関係数 τ は，

$$\tau = \frac{(A-B)}{n(n-1)/2}$$

によって得られる。ここで A は順位の大小関係が一致する組の数，B は不一致数である。
R コンソールで順位相関係数を計算するには，`cor.test()` 関数の中で，`method="spearman"`または `method="kendall"`と指定すれば良い。EZR では，「統計解析」「ノンパラメトリック検定」「相関係数の検定（Spearman の順位相関係数）」から，解析方法のところで Spearman か Kendall の横のラジオボタンを選んで OK ボタンをクリックすれば計算できる。

[a] ピアソンの相関係数の母相関係数を ρ と書き，スピアマンの順位相関係数を r_s と書く流儀もある。

1.3 回帰モデルの当てはめ

回帰は，従属変数のばらつきを独立変数のばらつきで説明するというモデルの当てはめである。十分な説明ができるモデルであれば，そのモデルに独立変数の値を代入することによって，対応する従属変数の値が予測あるいは推定できるし，従属変数の値を代入すると，対応する独立変数の値が逆算できる。こうした回帰モデルの実用例の最たるものが**検量線**である。検量線とは，実験において予め濃度がわかっている標準物質を測ったときの吸光度のばらつきが，その濃度によってほぼ完全に（通常 98% 以上）説明されるときに（そういう場合は，散布図を描くと，点々がだいたい直線上に乗るように見える），その関係を利用して，サンプルを測ったときの吸光度からサンプルの濃度を逆算するための回帰直線である（曲線の場合もあるが，通常は何らかの変換をほどこし，線形回帰にして利用する）。

検量線の計算には，(A) 試薬ブランクでゼロ点調整をした場合の原点を通る回帰直線を用いる場合と，(B) 純水でゼロ点調整をした場合の切片のある回帰直線を用いる場合がある。例えば，濃度の決まった標準希釈系列 (0, 1, 2, 5, 10 $\mu g/\ell$) について，純水でゼロ点調整をしたときの吸光度が，(0.24, 0.33, 0.54, 0.83, 1.32) だったとしよう。吸光度の変数を y，濃度を x と書けば，回帰モデルは $y = bx + a$ とおける。係数 a と b（a は切片，b は回帰係数と呼ばれる）は，次の偏差平方和を最小にするように，最小二乗法で推定される。

$$f(a,b) = \sum_{i=1}^{5} (y_i - bx_i - a)^2$$

この式を解くには，$f(a,b)$ を a ないし b で偏微分したものがゼロに等しいときを考えればいいので，次の２つの式が得られる。

$$b = \frac{\sum_{i=1}^{5} x_i y_i/5 - \sum_{i=1}^{5} x_i/5 \cdot \sum_{i=1}^{5} y_i/5}{\sum_{i=1}^{5} {x_i}^2/5 - \left(\sum_{i=1}^{5} x_i/5\right)^2}$$

$$a = \sum_{i=1}^{5} y_i/5 - b \cdot \sum_{i=1}^{5} x_i/5$$

これらの a と b の値と，未知の濃度のサンプルについて測定された吸光度（例えば 0.67 としよう）から，そのサンプルの濃度を求めることができる。注意すべきは，サンプルについて測定された吸光度が，標準希釈系列の吸光度の範囲内になければならないことである。回帰モデルが標準希釈系列の範囲外でも直線性を保っている保証は何もないのである[*4]。

R コンソールでは，`lm()`（linear model の略で線形モデルの意味）を使って，次のようにデータに当てはめた回帰モデルを得ることができる。

[*4] 回帰の外挿は薦められない。サンプルを希釈したり濃縮したりして吸光度を再測定し，標準希釈系列の範囲におさめることをお薦めする。

```
y <- c(0.24, 0.33, 0.54, 0.83, 1.32)
x <- c(0, 1, 2, 5, 10)
# 線形回帰モデルを当てはめる
res <- lm(y ~ x)
# 詳しい結果表示
summary(res)
# 散布図と回帰直線を表示する
plot(y ~ x)
abline(res)
# 吸光度 0.67 に対応する濃度を計算する
(0.67 - res$coef[1])/res$coef[2]
```

結果は次のように得られる。

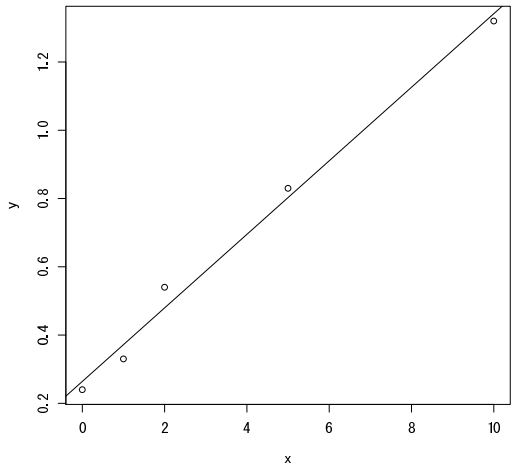
```
Call:
lm(formula = y ~ x)

Residuals:
       1        2        3        4        5
-0.02417 -0.04190  0.06037  0.02718 -0.02147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26417    0.03090   8.549 0.003363 **
x            0.10773    0.00606  17.776 0.000388 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04894 on 3 degrees of freedom
Multiple R-squared: 0.9906,     Adjusted R-squared: 0.9875
F-statistic:   316 on 1 and 3 DF,  p-value: 0.0003882
```

推定された切片は $a = 0.26417$，回帰係数は $b = 0.10773$ である。また，このモデルはデータの分散の 98.75% (0.9875) を説明していることが，`Adjusted R-squared` からわかる。また，`p-value` は，吸光度の分散がモデルによって説明される程度が誤差分散によって説明される程度と差が無いという帰無仮説の検定の有意確率である。



0.67 という吸光度に相当する濃度は，3.767084 となる。したがって，この溶液の濃度は，3.8 $\mu g/\ell$ だったと結論することができる。

EZR では，データはデータセットとして入力しなくてはいけない。「ファイル」「新しいデータセットを作成する」を選び，データセット名を入力：と書かれたテキストボックスに workingcurve と打って [OK] ボタンをクリックする。データエディタウィンドウが表示されたら，[var1] をクリックして，変数エディタの変数名というテキストボックスに y と打ち，型として "numeric" の方のラジオボタンをクリックしてから，キーボードの Enter キーを押す。次いで，同様にして [var2] を [x] に変える。それから，それぞれのセルに吸光度と濃度のデータを入力し，データエディタウィンドウを閉じる（通常は「ファイル」「閉じる」を選ぶ）。
散布図と回帰直線を描くには，「グラフ」「散布図」を選んで，x 変数として x を，y 変数として y を選び，[OK] ボタンをクリックする。
線形回帰モデルを当てはめるには，「統計解析」「連続変数の解析」「線形回帰（単回帰，重回帰）」と選び，目的変数として y，説明変数として x を選び，「モデル解析用に解析結果をアクティブモデルとして残す」の左のチェックボックスにチェックを入れて [OK] をクリックする。アウトプットウィンドウに結果が表示される（後述する多重共線性をチェックするために VIF を計算しようとしてエラーが表示されるが気にしなくて良い）。

検量線以外の状況でも，同じやり方で線形回帰モデルを当てはめることができる。survey データフレームに戻ってみよう[*5]。もし利き手の幅の分散を身長によって説明したいなら，線形回帰モデルを当てはめるには，R コンソールでは次のようにタイプすればいい。

```
res <- lm(Wr.Hnd ~ Height, data=survey)
summary(res)
```

EZR では，ロゴのすぐ右の "データセット:" の右側をクリックして survey を指定し，survey データセットをアクティブにしてから，「統計量」「モデルへの適合」「線形回帰」と選び，目的変数として Wr.Hnd，説明変数として Height を選び，「モデル解析用に解析結果をアクティブモデルとして残す」の左のチェックボックスにチェックを入れてから [OK] をクリックすると結果が得られる。

1.4 推定された係数の安定性を検定する

回帰直線のパラメータ（回帰係数 b と切片 a）の推定値の安定性を評価するためには，t 値が使われる。いま，Y と X の関係が $Y = a_0 + b_0X + e$ というモデルで表されるとして，誤差項 e が平均 0，分散 σ^2 の正規分布に従うものとすれば，切片の推定値 a も，平均 a_0，分散 $(\sigma^2/n)(1 + M^2/V)$（ただし M と V は x の平均と分散）の正規分布に従い，残差平方和 Q を誤差分散 σ^2 で割った Q/σ^2 が自由度 $(n - 2)$ のカイ二乗分布に従うことから，

$$t_0(a_0) = \frac{\sqrt{n(n-2)}(a - a_0)}{\sqrt{(1 + M^2/V)Q}}$$

が自由度 $(n - 2)$ の t 分布に従うことになる。

しかしこの値は a_0 がわからないと計算できない。a_0 が 0 に近ければこの式で $a_0 = 0$ と置いた値（つまり $t_0(0)$。これを切片に関する t 値と呼ぶ）を観測データから計算した値が $t_0(a_0)$ とほぼ一致し，自由度 $(n - 2)$ の t 分布に従うはずなので，その絶対値は 95% の確率で t 分布の 97.5% 点（サンプルサイズが大きければ約 2 である）よりも小さくなる。つまり，データから計算された t 値がそれより大きければ，切片は 0 でない可能性が高いことになるし，t 分布の分布関数を使えば，「切片が 0 である」という帰無仮説に対する有意確率が計算できる。

回帰係数についても同様に，

$$t_0(b) = \frac{\sqrt{n(n-2)V}b}{\sqrt{Q}}$$

が自由度 $(n - 2)$ の t 分布に従うことを利用して，「回帰係数が 0」であるという帰無仮説に対する有意確率が計算できる。有意確率が充分小さければ，切片や回帰係数がゼロでない何かの値をとるといえるので，これらの推定値は安定していることになる。

R コンソールでも EZR でも，線形回帰をした結果の中の，Pr(>|t|) というカラムに，これらの有意確率が示されている。

[*5] もちろん，survey データセットを使う前には，MASS パッケージをロードしておく必要がある。

2　回帰モデルの応用

2.1　重回帰モデル

説明変数は２つ以上の変数を含むことができる。このような場合，モデルは「重回帰モデル」と呼ばれる。注意しなくてはならない点がいくつかあるが，基本的には線形モデルの右側に + でつないで説明変数群を与えるだけである。

例えば，これまで扱ってきた survey データで，利き手の大きさの分散を説明するために，身長のみならず，利き手でない方の手の大きさも使うことにしよう。R コンソールでは次のように打てばよい（もちろん，予め MASS ライブラリをロードしておかねばならない）。

```
res <- lm(Wr.Hnd ~ Height + NW.Hnd, data=survey)
summary(res)
```

EZR では，まず「アクティブデータセット」の下の枠をクリックして survey を選び直してから，「統計解析」「連続変数の解析」「線形回帰（単回帰，重回帰）」を選び，“目的変数” として Wr.Hnd をクリックし，“説明変数” として Height をクリックしてからキーボードの Ctrl キーを押しながら NW.Hnd もクリックし，「モデル解析用に解析結果をアクティブモデルとして残す」の左のチェックボックスにチェックを入れてから [OK] ボタンをクリックすると，結果がアウトプットウィンドウに示される。

重回帰モデルでは，個々の説明変数について推定される回帰係数は，他の説明変数の目的変数への影響を調整した上で，その変数独自の目的変数への影響を示す「偏回帰係数」である。しかし偏回帰係数の値は，各変数の絶対的な大きさに依存しているので，各説明変数の目的変数への影響の相対的な強さを示すものにはならない。そうした比較をしたければ，R コンソールで次のようにタイプして stb として得られる「標準化偏回帰係数」が利用できる。結果をみると，Height の標準化偏回帰係数が 0.058，NW.Hnd の標準化偏回帰係数が 0.929 なので，利き手の大きさは大部分，利き手でない手の大きさによって説明されることがわかる。

```
sdd <- c(0,sd(res$model$Height),sd(res$model$NW.Hnd))
stb <- coef(res)*sdd/sd(res$model$Wr.Hnd)
stb
```

EZR には，メニューアイテムとしては，この機能は提供されていない。しかし，重回帰モデルを「残して」あるので，そのオブジェクトを使ったコマンドをスクリプトウィンドウに打つ。右上に「モデル：」とあるところの右側に，最後に生成されたモデル名が表示されている。これが RegModel.1 だとすると，上の 3 行の前に，res <- RegModel.1 という 1 行を追加すれば良い。その 4 行を選んでから，「実行」ボタンをクリックすれば，結果がアウトプットウィンドウに表示される。

2.2　当てはまりの良さの評価

データから得た回帰直線は，完璧にデータに乗ることはない。そこで，回帰直線の当てはまりのよさを評価する。a と b が決まったとして，$z_i = a + bx_i$ とおいたとき，$e_i = y_i - z_i$ を残差 (residual) と呼ぶ。残差は，y_i のばらつきのうち，回帰直線では説明できなかった残りに該当する。つまり，残差が大きいほど，回帰直線の当てはまりは悪いと考えられる。残差にはプラスもマイナスもあるので二乗和をとり，次の式で得られる「残差平方和」Q を定義することができる。

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - z_i)^2$$

$$= \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n - \frac{(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}/n$$

残差平方和 Q は回帰直線の当てはまりの悪さを示す尺度であり，それを n で割った Q/n を残差分散という。残差分散（var(e) と書くことにする）と Y の分散 var(Y) とピアソンの相関係数 r の間には，

$$\mathrm{var}(e) = \mathrm{var}(Y)(1 - r^2)$$

という関係が常に成り立つので，

$$r^2 = 1 - \mathrm{var}(e)/\mathrm{var}(Y)$$

となる。このことから r^2 が 1 に近いほど回帰直線の当てはまりがよいことになる。その意味で，r^2 を「決定係数」と呼ぶ。また，決定係数は，Y のばらつきがどの程度 X のばらつきによって説明されるかを意味するので，X の「寄与率」と呼ぶこともある。

データによっては，何通りもの回帰直線の残差平方和が大差ないという状況がありうる。例えば，目的変数と説明変数が実はまったく無関係であった場合は，データの重心を通るどのような傾きの線を引いても残差平方和はほとんど同じになってしまう。言い換えれば，傾きや切片の推定値が不安定になる。

r^2 は説明変数が多ければ大きくなるので，通常は自由度で r^2 を調整した「自由度調整済み重相関係数の二乗」を決定係数と考える。この値は，R コンソールでも EZR でも線形モデルの当てはめ結果の中で，`Adjusted R-Squared:` として表示されている。

当てはまりの良さの別の尺度として，AIC（赤池の情報量基準 : Akaike information criterion）も良く用いられる。とくに重回帰モデルでは，AIC も表示するのが普通である。R には `AIC()` という関数があり，線形回帰モデルの結果を付値したオブジェクトを，この関数に渡せば AIC が計算される（例えば `AIC(res)` のように使う）。ここでは AIC について詳しくは説明しないが，たくさんのオンライン資料や書籍で説明されている。

EZR で AIC を求めるには，標準化偏回帰係数を求めたときと同様に，スクリプトウィンドウに必要な関数を打ち，それを選択した上で「実行」ボタンをクリックする。モデルが `RegModel.1` であれば，必要な関数は，`AIC(RegModel.1)` である。

2.3 回帰モデルを当てはめる際の留意点

身長と体重のように，どちらも誤差を含んでいる可能性がある測定値である場合には，一方を説明変数，他方を目的変数とすることは妥当でないかもしれない（一般には，身長によって体重が決まるなど方向性が仮定できれば，身長を説明変数にしてもよいことになっている）。また，最小二乗推定の説明から自明なように，回帰式の両辺を入れ替えた回帰直線は一致しない。従って，どちらを目的変数とみなし，どちらを説明変数とみなすか，因果関係の方向性に基づいて（先行研究や臨床的知見を参照し）きちんと決めるべきである。

回帰を使って予測をするとき，外挿には注意が必要である。とくに検量線は外挿してはいけない。実際に測った濃度より濃かったり薄かったりするサンプルに対して，同じ関係が成り立つという保証はどこにもないからである（吸光度を y とする場合は，濃度が高くなると分子の重なりが増えるので飽和 (saturate) してしまい，吸光度の相対的な上がり方が小さくなっていき，直線から外れていく）。サンプルを希釈したり濃縮したりして，検量線の範囲内で定量しなくてはいけない。

例題

組み込みデータ `airquality` は，1973 年 5 月 1 日から 9 月 30 日まで 154 日間のニューヨーク市の大気環境データである。含まれている変数は，`Ozone`（ppb 単位でのオゾン濃度），`Solar.R`（セントラルパークでの 8:00 から 12:00 までの 4000 から 7700 オングストロームの周波数帯の太陽放射の強さを Langley 単位で表した値），`Wind`（LaGuardia 空港での 7:00 から 10:00 までの平均風速，マイル／時），`Temp`（華氏での日最高気温），`Month`（月），`Day`（日）である。日照の強さを説明変数，オゾン濃度を目的変数として回帰分析せよ。

R コンソールでは，次の 4 行を打てば良い。

```
plot(Ozone ~ Solar.R, data=airquality)
res <- lm(Ozone ~ Solar.R, data=airquality)
abline(res)
summary(res)
```

EZR では，まず「ファイル」「パッケージに含まれるデータを読み込む」から左の枠の `datasets` をダブルクリックし，右の枠に現れるデータフーレムの下の方へスクロールして `airquality` をダブルクリックしてから OK ボタンをクリックして `airquality` データフレームをアクティブにする。次いで「グラフ」「散布図」を選び，x 変数を `Solar.R`，y 変数を `Ozone` として [OK] をクリックする。次に，「統計解析」「連続変数の解析」「線形回帰」を選ぶ。目的変数として `Ozone` を，説明変数として `Solar.R` を選んで OK ボタンをクリックする。

R コンソールでも EZR でも得られる結果は同じで，次の枠内の通りである。

```
Call:
lm(formula = Ozone ~ Solar.R, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-48.292 -21.361  -8.864  16.373 119.136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.59873    6.74790   2.756 0.006856 **
Solar.R      0.12717    0.03278   3.880 0.000179 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 31.33 on 109 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-Squared: 0.1213,Adjusted R-squared: 0.1133
F-statistic: 15.05 on 1 and 109 DF,  p-value: 0.0001793
```

得られた回帰式は Ozone = 18.599 + 0.127 · Solar.R であり，最下行をみると F 検定の結果の p 値が 0.0001793 ときわめて小さいので，モデルの当てはまりは有意である。しかし，その上の行の `Adjusted R-squared` の値が 0.11 ということは，このモデルではオゾン濃度のばらつきの 10% 余りしか説明されないことになり，あまりいい回帰モデルではない。

当てはまりを改善するには，説明変数を追加することが有効な場合がある。この例では，`Wind` あるいは `Temp` を説明変数に加えて重回帰モデルにすれば，当てはまりが改善する。R コンソールでは，次の 3 行を打てば重回帰モデルの当てはめができる。自由度調整済み重相関係数の二乗が約 60% にまで改善していることがわかる。

```
mres <- lm(Ozone ~ Solar.R + Wind + Temp, data=airquality)
summary(mres)
AIC(mres)
```

EZR では，「統計解析」「連続変数の解析」「線形回帰」を選ぶ。"目的変数" として Ozone を選び，"説明変数" として Solar.R をクリックしてからキーボードの Ctrl キーを押しながら Wind と Temp もクリックし，「モデル解析用に解析結果をアクティブモデルとして残す」の左のチェックボックスにチェックを入れてから [OK] ボタンをクリックすると，同じ結果が得られる。

3 文献

- 青木繁伸 (2009) R による統計解析. オーム社
- 大橋靖雄，浜田知久馬 (1995) 生存時間解析: SAS による生物統計. 東京大学出版会.
- 古川俊之 [監修]，丹後俊郎 [著] (1983) 医学への統計学. 朝倉書店.
- 中澤　港 (2003) R による統計解析の基礎. ピアソン・エデュケーション.
- 中澤　港 (2007) R による保健医療データ解析演習. ピアソン・エデュケーション.
- 永田　靖 (2003) サンプルサイズの決め方. 朝倉書店.
- http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf（作成者である John Fox 自身による R Commander の入門テキスト）
- http://www.ec.kansai-u.ac.jp/user/arakit/documents/Getting-Started-with-the-Rcmdr-ja.pdf（日本語版メニュー作成者である荒木孝治さんによる邦訳）
- 神田善伸 (2012) EZR でやさしく学ぶ統計学〜EBM の実践から臨床研究まで〜，中外医学社[*6].

[*6] http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html