医学情報処理演習第８回「相関と回帰」[*1]

2006 年 11 月 27 日 中澤　港 (nminato@med.gunma-u.ac.jp)

前回まで２回は平均値の差の検定を行ったが，今回は２つの変数間の関係を扱う。

相関と回帰の違い

相関と回帰は，どちらも２つの変数の関係を扱うので混同されやすいが思想は異なる。相関は，変数間の関連の強さを表すものである。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ものである。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。

相関

関係とか関連とかいっても，その中身は多様である。例えば，$pV = nRT$ のような物理法則は，測定誤差を別にすれば 100%成り立つ関係である。身長と体重の間には，そのような例外のない関係は成り立たない（つまり，**関係にばらつきがある**）。しかし，無関係ではないことは直感的にも理解できるし，「身長の高い人は体重も概して重い傾向がある」ことは間違いない。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。

見かけの相関・擬似相関

相関関係があっても，それが見かけ上のものである（それらの変量がともに，別の変量と真の相関関係をもっている）場合がある。具体例としては，血圧と所得の間に正の相関があるという命題は，データをとってみれば，多くの場合に成り立つであろう。しかしこれは，おそらくどちらも年齢や摂取エネルギー量との間に真の相関関係が存在するのであって，それらの影響を制御したら（例えば同年齢で同じような食生活をしている人だけについて見る，という層別化をしたら），血圧と所得の間の正の相関は消えてしまうだろう。この場合，見かけ上の相関があることは，たまたまそのデータで成り立っているだけであって，科学的仮説としての意味に乏しい。

時系列データや地域相関のデータでは，擬似相関 (spurious correalation) が見られる場合もある。例えば，日本の砂糖輸入量と溺死・溺水者数の年次別データをプロットしてみると，負の相関関係があるように見えるのだが，両者の間には真の関係はない。ある年に日本で植えた木の幹の太さと，同じ年にイギリスで生れた少年の身長を 15 年分，毎年１回測ったデータをプロットすると，おそらくは正の相関関係があるように見えるのだが，両者の間には関係がないのは明らかである（どちらも年次と真の相関があるとはいえるだろうが）。複数の種類の異なるデータをまとめて見ることで見かけの相関が生じてしまう場合もあるので，注意が必要である。

線形の相関・非線形の相関

上で定義したように，相関関係は増減をともにする関係であればいいので，その関係が線形（一次式で表される，散布図で直線として表される）であろうと非線形（二次以上の多項式または階段関数などで表される）であろうと問題ない。しかし，一般には，線形の関係があるという限定的な意味で使われる場合が多い。なぜ

[*1] 本資料は http://phi.med.gunma-u.ac.jp/medstat/it08-2006.pdf としてダウンロード可能である。

なら，相関を表すための代表的な指標である相関係数[*2]r が，線形の関係の強さを示すための指標だからである。より厳密にいえば，r が「線形の関係の強さを示す指標」であるためには，その 2 つの変数が二次元正規分布に従っていることを前提とする。

非線形の相関関係を捉えるには，2 つのアプローチがある。1 つは線形になるように対数変換などの変換をほどこすことで，もう 1 つはノンパラメトリックな相関係数（分布の形によらない，例えば順位の情報だけを使った相関係数）を使うことである。ノンパラメトリックな相関係数にはスピアマン (Spearman) の順位相関係数 ρ や，ケンドール (Kendall) の順位相関係数 τ がある。

ピアソンの積率相関係数とは，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る。式で書けば，相関係数の推定値 r は，X の平均を $\bar{X}$，Y の平均を $\bar{Y}$ と書けば，

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

となる。母相関係数がゼロかどうかという両側検定のためには，それがゼロであるという帰無仮説の下で，**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検定すればよい[*3]。

R で変数 `X` と `Y` の相関係数を計算して有意確率を得るには下枠内の 4 行を打てばよいが，実はそれと同じことが `cor.test(X,Y)` でできてしまう（そればかりでなく，信頼区間も計算してくれる）。1 行目の相関係数を計算する部分も `print(r <- cor(X,Y))` で置き換え可能である。

```
print(r <- cov(X,Y)/sqrt(var(X)*var(Y)))
n <- length(X)
t0 <- r*sqrt(n-2)/sqrt(1-r^2)
print(2*(1-pt(abs(t0),n-2)))
```

信頼区間は，サンプルサイズがある程度大きければ（通常は 20 以上），正規近似を使って計算できる。すなわち，

$$a = \frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{\sqrt{n-3}}Z(\alpha/2), \quad b = \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{1}{\sqrt{n-3}}Z(\alpha/2)$$

と書くことにすると[*4]，母相関係数の $100\times(1-\alpha)\%$信頼区間の下限は $(\exp(2a)-1)/(\exp(2a)+1)$，上限は $(\exp(2b)-1)/(\exp(2b)+1)$ である[*5]。

順位相関係数は，非線形の相関関係を捉えたい場合以外にも，分布が歪んでいたり，外れ値がある場合に使うと有効である。スピアマンの順位相関係数 ρ は[*6]，値を順位で置き換えた（同順位には平均順位を与えた）

[*2] 普通，ただ相関係数といえば，ピアソンの積率相関係数 (Pearson's Product Moment Correlation Coefficient) を指し，通常，r という記号で表す。

[*3] 繰り返すが，検定は，帰無仮説を立てて，それが正しいときに，現在得られているデータ以上に外れたデータが偶然得られる確率がどれほどかを計算し（有意確率），その確率が 0.05 とか 0.01 といった有意水準より小さいときに，偶然ではありえないほど小さいと判断し，帰無仮説を棄却するという意思決定を行うためのプロセスである。有意確率を計算するためには，通常，帰無仮説が正しいとしたときに既知の確率分布に従うはずの量（**検定統計量**）を計算し，その既知の確率分布の分布関数のその値に対応する値を 1 から引けば（片側検定のとき）有意確率となる。両側検定の場合はその確率を 2 倍する。この原理は，たいていの検定に共通している。

[*4] $Z(\alpha/2)$ は標準正規分布の $100\times(1-\alpha/2)$ パーセント点である。α を `alpha` と書けば `qnorm(1-alpha/2,0,1)` で得られる。例えば有意水準 5%，すなわち $\alpha = 0.05$ なら，`qnorm(0.975,0,1)` とする。

[*5] なお，ln は自然対数，exp は指数関数を表す。この式から明らかなように，母相関係数の信頼区間はどんなに広がっても下限は -1 以下にはならず，上限は 1 以上にならない。

[*6] 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。

ピアソンの積率相関係数になる。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かどうかという両側検定は，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 では `cor.test(X, Y, method="pearson")` とすれば（あるいは `method=`オプションをつけなければ）ピアソンの相関係数が，`cor.test(X, Y, method="spearman")` でスピアマンの順位相関係数が，`cor.test(X, Y, method="kendall")` でケンドールの順位相関係数が得られる。同時に，`alternative` を指定しないときは，「相関係数がゼロである」を帰無仮説として両側検定した有意確率と 95%信頼区間が表示される。なお，例えば `cor.test(X, Y, alternative="g")` とすれば，ピアソンの相関係数が計算され，対立仮説を「正の相関がある」とした片側検定の結果が得られる。なお，ケンドールに関しては並べ換えによる正確な確率も求めることができ，その場合は `exact=T` というオプションを指定する。

例題

`ToothGrowth` は各群 10 匹ずつのモルモットに 3 段階の用量のビタミン C をアスコルビン酸としてあるいはオレンジジュースとして投与したときの象牙芽細胞（歯）の長さを比較するデータである。変数 `len` が長さ，`supp` が投与方法，`dose` が用量を示す。投与方法の違いを無視して用量と長さの相関関係を調べよ。

まず，`attach(ToothGrowth)` して `ToothGrowth` に含まれている変数が使えるようにする。作図によって用量と長さの関係を概観するためには，散布図を描けばよいので，横軸を `dose`，縦軸を `len` としたプロットをするために，`plot(dose,len)` とする。なんとなく `dose` が増すにつれて `len` が長くなっていくような，正の相関関係があるように見えるだろう。そこで，この相関係数を算出し，「相関係数がゼロと差がない」という帰無仮説を検定してみるために，`cor.test(dose,len)` と打てば，以下の出力が得られる。

```
        Pearson's product-moment correlation

data:  dose and len
t = 10.2501, df = 58, p-value = 1.243e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6892521 0.8777169
sample estimates:
      cor
0.8026913
```

この結果をみると，ピアソンの相関係数の推定値は 0.80，95%信頼区間が [0.69, 0.88]（ここでは四捨五入で示しているが，真の区間を含む最小の幅に丸める方がよいという意見もあり，それなら [0.68, 0.88] と記載する）となる。95%信頼区間がゼロを含んでいないので，**帰無仮説が有意水準 5%で棄却されるの**は明らかだが，有意確率をみても，`p-value=1.243e-14` とほとんどゼロであることが確認できる。

練習

`method="spearman"`とか `method="kendall"`でも試してみよう。

回帰

実験によって，あるサンプルの濃度を求めるやり方の１つに，検量線の利用がある。検量線とは，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。

いずれも，量がわかっている**「独立変数」**（この場合は濃度）を x，誤差を含んでいる可能性がある測定値である**「従属変数」**（この場合は吸光度）を y として $y = bx + a$ という形の回帰式の係数 a と b を最小二乗法で推定し，**サンプルを測定した値 y から $x = (y - a)/b$ によってサンプルの濃度 x を求める**。測定値から濃度を推定するときには，回帰式をそのまま使うのではなく，**逆算する**形になるので注意が必要である。

回帰直線の適合度の目安としては，学生実習でも相関係数の 2 乗が 0.98 以上あることが望ましい。また，データ点の最小，最大より外で直線関係が成立する保証はない。従って，サンプル測定値が標準物質の測定値の最小より低いか，最大より高いときは，限界を超えていることになってしまう[*7]。

測定点 $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ が得られたときに，検量線 $y = bx + a$ を推定するには，図に示した線分の二乗和が最小になるように a と b を設定すればよい，というのが最小二乗法の考え方である[*8]。つまり，

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

$$= b^2 \sum_{i=1}^{n} x_i^2 - 2b \sum_{i=1}^{n} x_i y_i + 2ab \sum_{i=1}^{n} x_i - 2a \sum_{i=1}^{n} y_i + na^2 + \sum_{i=1}^{n} y_i^2$$

が最小になるような a と b を推定すればよい。通常，a と b で偏微分した値がそれぞれ 0 となることを利用して計算すると簡単である。つまり，

$$\frac{\partial f(a,b)}{\partial a} = 2na + 2(b \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i) = 0$$

$$i.e. \quad na = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i$$

$$i.e. \quad a = (y \text{ の平均}) - (x \text{ の平均}) * b$$

$$\frac{\partial f(a,b)}{\partial b} = 2b \sum_{i=1}^{n} x_i^2 + 2(a \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i) = 0$$

$$i.e. \quad b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i$$

を連立方程式として a と b について解けばよい。これを解くと，

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

[*7] 余談だが，このような場合はサンプルを希釈するか濃縮して測定するのが普通である。

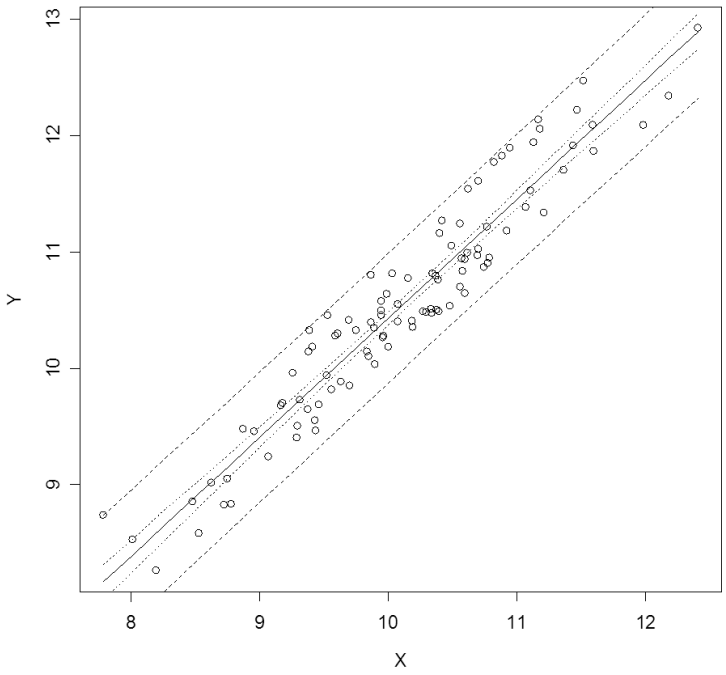
[*8] なお，試薬ブランクでゼロ点調整をした場合の原点を通る回帰直線を検量線とするには，$y = bx$ について同じ手順で計算すればよいので，$b = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} {x_i}^2$ となる。これを実行するための R のコードは `lm(Y~X-1)` または `lm(Y~0+X)` である。

が得られる[*9]。b の値を上の式に代入すれば a も得られる。**検量線に限らず，一般の回帰直線でも，計算方法は原則として同じ**である。名称の説明をしておくと，**一般に，$y = bx + a$ という回帰直線について，b を回帰係数 (regression coefficient)，a を切片 (intercept) と呼ぶ。**

R では，線形回帰を行うための関数は `lm()` である。例えば，`lm(Y~X)` のように用いれば，回帰直線の推定値が得られる。散布図の上に回帰直線を重ね描きするには，`plot(Y~X)` としてから，`abline(lm(Y~X))` とすればよい。また，回帰直線の 95%予測区間（期待値の標準誤差が 95%の信頼性をもってこの区間に含まれるであろう範囲）と 95%信頼区間（データの 95%がこの範囲に入るであろうという範囲）を点線と破線で重ね描きするには，`predict()` 関数を使って範囲を予測し，重ね描きすればよい[*10]。例えば，下の例のようになる。

it08-1-2006.R

```
RNGkind("Mersenne-Twister")
set.seed(1)
X <- rnorm(100,10,1)
Y <- X + runif(100,0,1)
summary(res <- lm(Y~X))
XX <- data.frame(X=seq(min(X),max(X),length=20))
plim <- predict(res, XX, interval="prediction")
clim <- predict(res, XX, interval="confidence")
plot(X,Y)
matlines(XX,plim,col=1,lty=c(1,2,2))
matlines(XX,clim,col=1,lty=c(1,3,3))
```



5 行目で回帰分析を実行した結果を `res` というオブジェクトに付値するのと同時に，決定係数や回帰係数と切片の検定結果を出力している。出力結果は次の枠内の通りである。

[*9] 分母分子を n^2 で割れば，b は $x_i y_i$ の平均から x_i の平均と y_i の平均の積を引いて，x_i の二乗の平均から x_i の平均の二乗を引いた値で割った形になる。

[*10] ただし `predict()` 関数は回帰式の計算値そのものも返すので，これで重ね描きする場合は `abline()` は不要である。

Residuals:の部分は残差を示す。残差とは，回帰による予測値と実測値の差である。独立変数（ここではX）の最小値，第1四分位，中央値，第3四分位，最大値に対応する従属変数の値から，回帰式にそれらの独立変数の値を代入して得られる値（これが回帰による予測値）を引いた値を意味する。これが0に近いほど回帰式のデータへの当てはまりは良いと考えられる。

次のCoefficients:のところに表示されるのが，さまざまな係数とその検定結果である。(Intercept)の行は切片を示す。Xの行が変数Xについての情報を与える。Estimateの値が切片と回帰係数の点推定量であり，Std. Errorの列はそれぞれの標準誤差を示す。t valueは，「切片がゼロと差がない」及び「回帰係数がゼロと差がない」を帰無仮説とする検定を行うための，t分布に従う検定統計量である。Pr(>|t|)は有意確率を示す。下の方に，Adjusted R-squaredとあるのが自由度調整済み相関係数の二乗で，後述するように決定係数とも呼ばれ，従属変数Yのばらつきのどれくらいの割合が独立変数Xのばらつきによって説明されるかを示す値である。このデータでは92%近く，説明力の強い回帰式が得られたといえる。

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39958 -0.24095 -0.04863  0.20490  0.57297

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17392    0.31703   0.549    0.585
X            1.02583    0.03124  32.837   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2792 on 98 degrees of freedom
Multiple R-Squared: 0.9167, Adjusted R-squared: 0.9158
F-statistic:  1078 on 1 and 98 DF,  p-value: < 2.2e-16
```

検量線の例題

血清鉄濃度をFe-Test Wakoというキットで測定するため，鉄の標準希釈系列を0, 0.5, 1, 2 (mg/L)として作成し，それをこのキットで処理して発色させた溶液の波長562 nmの吸光度を測った結果が，0.012, 0.058, 0.104, 0.193として得られた。これから検量線を求めて，測定に使えるかどうか評価せよ。次に，6人の血清サンプルを同じ方法で処理して発色させた溶液の吸光度が0.107, 0.075, 0.077, 0.099, 0.096, 0.108だったときに，この6人の血清鉄濃度を求めよ。

まず濃度をconc，吸光度をabsとしてデータを入力し，concを横軸，absを縦軸にして散布図を描く[*11]。

```
conc <- c(0, 0.5, 1, 2)
abs <- c(0.012, 0.058, 0.104, 0.193)
plot(abs~conc)
```

だいたい直線に乗っているように見えるので，回帰分析をしてみる。以下のように，回帰分析の結果をいったんresに保存しておくと，描画や表示やその後の計算に便利である。

[*11] 描画命令は，ここではplot(abs~conc)としたが，plot(conc,abs)でも同じことである。

```
res <- lm(abs~conc)
abline(res)
summary(res)
```

出力される結果から，Adjusted R-squared: 0.9999 なので検量線として使ってよいと判断できる。回帰係数は 0.0904571，切片は 0.0126000 として得られているが，これらの値はそれぞれ res$coef[2]，res$coef[1] として参照できるので[*12]，サンプルの吸光度から濃度を**逆算**するときには，下枠内のように変数名のまま参照した方が間違えない[*13]。

```
dat <- c(0.107, 0.075, 0.077, 0.099, 0.096, 0.108)
(dat-res$coef[1])/res$coef[2]
```

決定係数

データから得た回帰直線は，$pV = nRT$ のような物理法則と違って，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必要が出てくる。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，

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

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

とすれば，Q は，回帰直線の当てはまりの悪さを示す尺度となる。この Q を「残差平方和」と呼び，それを n で割った Q/n を残差分散 $var(e)$ という。残差分散 $var(e)$ と Y の分散 $var(Y)$ とピアソンの相関係数 r の間には，$var(e) = var(Y)(1 - r^2)$ という関係が常に成り立つので，$r^2 = 1 - var(e)/var(Y)$ 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 Y への**「寄与率」**と呼ぶこともある。

なお，モデルのデータへの当てはまりを評価する指標は，残差分散と決定係数の他にも，AIC や BIC や Deviance などいろいろある。一般化線型モデルのところで詳しく触れる。

回帰直線推定と検定のしくみ

回帰直線は最小二乗法でもっとも残差平方和が小さくなるように選ぶわけだが，データの配置によっては，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，独立変数と従属変数（として選んだ変数）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。その意味で，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ことが大事である。そのためには，t 値というものが使われている。いま，Y と X の関係が $Y = a_0 + b_0 X + e$ 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

[*12] res$coef の部分は coef(res) でも同じ意味である。

[*13] http://phi.med.gunma-u.ac.jp/medstat/it08-2-2006.R で，この一連の過程がダウンロードできる。

ば，回帰係数の推定値 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。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ことになる。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。既に示したとおり，これらの検定結果は `summary(lm())` で表示される。

独立変数・従属変数と因果の向き

実は，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，どちらかを独立変数，どちらかを従属変数，とみなすことに問題がある。一般には，身長によって体重が決まってくるというように方向性が仮定できれば，身長を独立変数と見なしてもよいことになっているが，回帰分析をしてしまうと，独立変数に測定誤差がある可能性が排除されてしまうことには注意しておくべきである。つまり，測定誤差が大きい可能性がある変数を独立変数とした回帰分析は，できれば避けたほうが良い[*14]。また，最小二乗推定の説明から自明なように，独立変数と従属変数を入れ替えた回帰直線は一致しないので，どちらを従属変数とみなし，どちらを独立変数とみなすか，因果関係の方向性に基づいてきちんと決めねばならない。

回帰式を予測に用いる際の留意点

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。前述の通り，検量線は，原則として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）。しかし，外挿による予測は，実際にはかなり行われている。例えば世界人口の将来予測とか，河川工学における基本高水計算式とか，感染症の発症数の将来予測は，回帰の外挿による場合が多い（**この場合は逆算ではなく，実際にデータを得られていない横軸の値を代入したときに縦軸の値がいくつになるかを予測することになる**）。このやり方が妥当性をもつためには，その回帰関係が（1）かなり説明力が大きく，（2）因果関係がある程度認められ，（3）それぞれの変数の分布が端の切れた分布でない（truncated distribution でない）という条件を満たす必要がある。そうでない場合は，その予測結果が正しい保証はどこにもない。

[*14] そうもいかないのが実情だが。

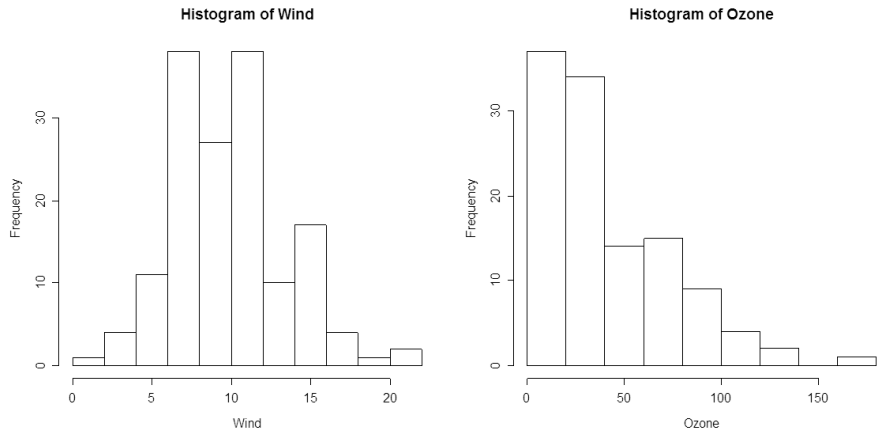
例題

`data(airquality)` とすると，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を使うことができる。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。
このデータから，オゾン濃度と風速の関係について検討し，もし風速からオゾン濃度を予想できるとしたら，風速 15 マイル／時の日のオゾン濃度はどうなるか，また，もしも 25 マイル／時の日があったとしたらどうなるか，期待値とその 95%信頼区間（標準誤差から計算される，母集団における期待値の信頼区間）を計算せよ。

データフレーム `airquality` を `attach` し，まず風速とオゾン濃度の分布の正規性について Shapiro-Wilk の検定をし，ヒストグラムによってそれぞれの分布の様子をみてみる。

it08-3-2006.R

```
attach(airquality)
shapiro.test(Wind)
shapiro.test(Ozone)
layout(t(1:2))
hist(Wind)
hist(Ozone)
```



風速は正規分布に従っているがオゾン濃度は正規分布に従っているといえないので，

```
cOzone <- log(Ozone+10)
```

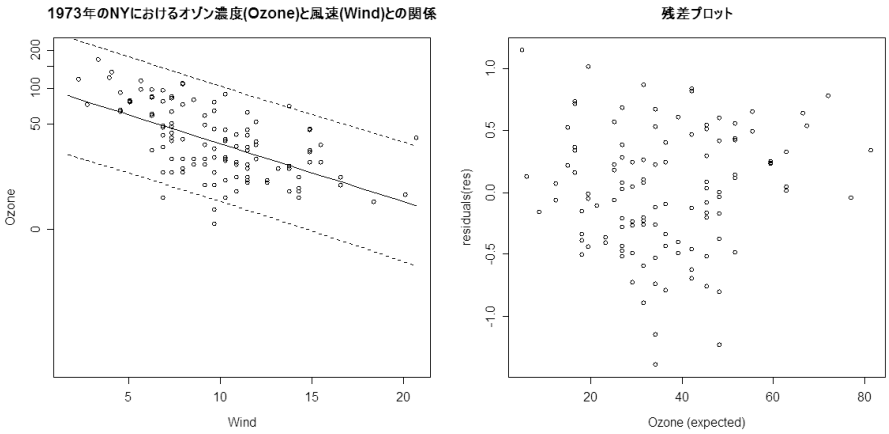
と変数変換する[*15]。そこで，次に大雑把に風速とオゾン濃度の関係をみるために散布図をプロットしてみる。

散布図をみると，確かに風速が大きくなるほどオゾン濃度が下がる関係がありそうにみえる。そこで，独立変数を `Wind`，従属変数を `cOzone` とする回帰分析をしてみる。95%予測区間（データの 95%がそこに入るであろう範囲）付きで，回帰直線をさっきの散布図に重ね描きしてみると，風速が中程度のところで回帰直線よりオゾン濃度が低い方に外れた値がいくつかあり，回帰関係全体がそれによって歪んでいる可能性が考えられる。したがって，この図から厳しく判断するなら，この回帰関係は予測に使うべきではない。残差プロット

[*15] このデータの場合，ただの対数変換では正規分布に従うという帰無仮説はまだ棄却されるので，別の変換が必要となる。例えば，10 を加えた後で対数変換するとか，または立方根変換すれば，正規分布に従うという帰無仮説が棄却されなくなる。メカニズムを考えると，風速が無限大に飛べばオゾン濃度は限りなくゼロに近づきそうだが，風速ゼロのときでもオゾン濃度が無限大になることはなさそうなので，10 を加えて対数変換してみた（これが最適とは限らない）。なお，変換するかどうかは，正規性の検定で有意なら必ずするわけではなく，変換によるひずみと正規分布に従わないことによるひずみを勘案して決定する。

（右側）をみても，それほど残差は大きくないものの，やはり中央付近が凹んでいるようにみえる[*16]。

```
layout(t(1:2))
cOzone <- log(Ozone+10)
shapiro.test(cOzone)
plot(cOzone~Wind,yaxt="n",ylab="Ozone",ylim=c(0,log(210)),
 main="1973 年の NY におけるオゾン濃度 (Ozone) と風速 (Wind) との関係")
yi <- 0:4*50
axis(2,log(yi+10),yi)
res <- lm(cOzone~Wind)
X <- data.frame(Wind=seq(min(Wind),max(Wind),length=20))
Y <- predict(res,X,interval="predict")
matlines(X,Y,col=1,lty=c(1,2,2))
plot(exp(fitted.values(res))-10,residuals(res),main="残差プロット",xlab="Ozone (expected)")
exp(predict(res,list(Wind=15),interval="confidence"))-10
exp(predict(res,list(Wind=25),interval="confidence"))-10
summary(res)
detach(airquality)
```



そうはいっても，この程度のズレなら，データのある範囲内なら回帰式を使えないこともない。回帰式は

$$\log(\mathrm{Ozone}+10) = 4.74 - 0.0985 \cdot \mathrm{Wind}$$

であり，「回帰係数がゼロと差がない」帰無仮説の検定の有意確率は 6.8×10^{-12} なので帰無仮説は棄却されるし，決定係数は 0.33 なので，オゾン濃度のばらつきの 33%は風速のばらつきによって説明されると考えられる（繰り返すが，これは予測には十分ではない）。

それでも強引に `Wind` が 15（マイル／時）のときのオゾン濃度の期待値と期待値の 95%信頼区間を求めたところ，$16.2[12.3, 20.7]$ (ppb) であった。

しかし `Wind` が 25 のときの期待値と 95%信頼区間は $-0.2[-3.4, 4.5]$ (ppb) となり，期待値さえ負の値というありえない結果になってしまう。これは (1) 回帰式のデータへの当てはまりが不十分かつ (2) データのない範囲への外挿なので式が成り立つ保障がそもそもないという 2 点を考えれば当然の結果である。したがって，風速 25 マイル／時の日の予測は不可能といえる。

[*16] 残差プロットの横軸は `fitted.values(res)` だと回帰式による従属変数の推定値そのものになるが，ここでは変数変換しているので，それを元に戻す（逆変換）ために `exp(fitted.values(res))-10` とした。横軸は `res$model$X` として独立変数の値をとる場合もある。また，残差プロットの縦軸 `residuals(res)` は，個々のデータと従属変数の推定値の差になり，これも変数変換後の値だが，対数をとって差を出すと，比の対数になるので，指数をとって元に戻しても残差にならない（残比とでもいうべきか）ため，逆変換していない。

課題

`http://phi.med.gunma-u.ac.jp/medstat/p08.txt` は，ソロモン諸島のある村に居住する成人女性 17 人の身体計測データで，含まれている変数は身長 (`HT`，単位は cm)，体重 (`WT`，単位は kg)，Body Mass Index(`BMI`，単位は kg/m^2)，タニタの体脂肪計つき体重計で測定した体脂肪割合 (`FAT`，単位は%)，収縮期血圧 (`SBP`，単位は mmHg) である。なお，言うまでもないが，BMI は身長と体重からの計算値である。

このデータから，以下のどちらかに答えよ（余力があれば両方でもよい）。

(1) BMI と体脂肪割合の相関について検討せよ。

(2) 身長を独立変数，体重を従属変数とした回帰分析を行って，もし次に測定した人の身長が 155 cm だったら，その人の体重は何 kg と予想されるか，95%信頼区間をつけて推定せよ。

結果は，作図したものをパワーポイントに貼り付け，推定や検定の結果をテキストボックスで記入し，学籍番号，氏名も記入した上でプリントアウトし，**氏名を自筆して**提出すること。結果の提出をもって出席確認とする。