医学基礎技術演習・実験基本技術（医学統計学）テキスト (2)

中澤　港（公衆衛生学　准教授）nminato@med.gunma-u.ac.jp

2008 年 6 月 4 日

1　相関と回帰

相関も回帰も 2 つの量的な変数間の関係を調べる点は共通である。そのため，まずは散布図を描く。

例題 1

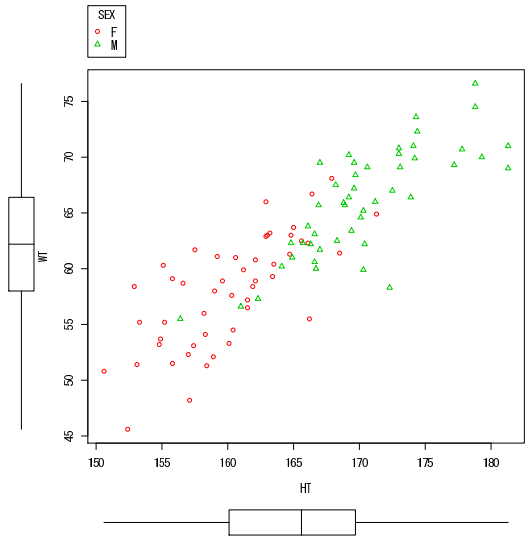
`http://phi.med.gunma-u.ac.jp/msb/data/p01.txt` は，男女合わせて 100 人の集団の身長 (HT) と体重 (WT) のデータ（欠損値を含む）である。身長を横軸，体重を縦軸とした散布図を描け。記号は性別 (SEX) ごとに変えること。

まず，先週やったように R のアイコンをダブルクリックして R を起動後，プロンプトに `library(Rcmdr)` として Rcmdr を起動する。

次いでデータを読み込む。Rcmdr のメニューの**「データ」から「データのインポート」の「テキストファイルまたはクリップボードから」**を選んで，表示されるダイアログで「区切り」を「タブ」に変え，OK ボタンをクリックして，ファイルを選択するウィンドウが出てきたら，ファイル名を入力する枠にデータの URL を打って OK するとネットワーク経由でデータファイルをデータフレームとして読み込むことができる。とくに変えていなければ，データフレーム名は `Dataset` となっているはずである（ここまでは前回の復習）。

次いで，散布図を描く。**Rcmdr のメニューの「グラフ」から「散布図」を選ぶ。**表示されるウィンドウの中で，「x 変数」として HT を選び，「y 変数」として WT を選ぶ。下の方は，周辺箱ヒゲ図と最小 2 乗直線と平滑線の右側のボックスにチェックが入っているが，相関をみる場合は最小 2 乗直線のチェックは外す（平滑線もない方がいい）。周辺箱ヒゲ図は横軸の変数，縦軸の変数別々に箱ヒゲ図を描いてくれるので，チェックが入ったままでよい。

この例題では性別にプロット記号を変えることとなっているので，下の方の「層別のプロット」というボタンをクリックして，出てくるウィンドウの中で層別変数として SEX を選ぶ。その下の層別して線を描くというボックスにチェックが入っているが，最小 2 乗直線のチェックを外してあれば，このボックスの指定は無効である。後は OK ボタンをクリックしていけば，次の散布図ができる。



1.1 相関と回帰の違い

相関と回帰は混同されやすいが，思想はまったく違う。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。

1.2 相関関係とは

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。身長と体重の間の関係はそうではないが，無関係ではないことは直感的にも理解できるし，散布図を見ても「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。

1.3 見かけの相関，擬似相関

相関関係があっても，それが見かけ上の関係に過ぎない場合がある。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，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，という限定をしたら），相関関係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，両者の間に直接関係がないのは明らかである（どちらも時間が経つにつれて大きくなっているだけである）。この場合でなくても，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こともあるので，注意が必要である。

1.4 直線的な相関，直線に載らない相関

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，順位の情報だけを使った相関の指標（順位相関係数）を計算する。

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，r という記号で表すが，この r は直線的な関係の強さの指標である。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，

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

である。

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）がどれほど小さいかを調べ，例えば 5% 未満ならば，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。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る。

例題 2

例題 1 のデータで身長と体重のピアソンの積率相関係数を計算し，有意性を検定せよ。

Rcmdr では，「統計量」の「要約」の「相関の検定」を選び，変数として WT と HT 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）。相関のタイプとして「ピアソンの積率相関」と「スピアマンの順位」と「ケンドールのタウ」が選べるようになっている。この例題ではピアソンの積率相関係数を求めるので，初期設定のまま「ピアソンの積率相関」にしておけばよい。検定についても「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:  Dataset$HT and Dataset$WT
t = 16.4519, df = 95, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7977988 0.9045751
sample estimates:
     cor
0.860348
```

これより，身長と体重の関係について求めたピアソンの積率相関係数は，$r = 0.86$（95% 信頼区間が $[0.798, 0.905]$）であり，`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」とみなすのが目安である。

せっかく男女別にプロットしたので，相関係数の検定も男女別に実行したいところだが，残念ながらボタン 1 つというわけにはいかない。男女別に相関係数の検定を実行するには，**「データ」の「アクティブデータセット」の「アクティブデータセットの部分集合を抽出」**を使って男女別のデータフレームを作成しなくてはならない。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でよく,「部分集合の表現」のボックスに SEX=="M"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「統計量」「要約」「相関の検定」をすれば男性の身長と体重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。

```
Pearson's product-moment correlation

data:  Males$HT and Males$WT
t = 8.636, df = 46, p-value = 3.476e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6468709 0.8750522
sample estimates:
      cor
0.7864562
```

女性について同じことをするには，まず**「データ」の「アクティブデータセット」の「アクティブデータセットの選択」**で Dataset を選び直し,「アクティブデータセットの部分集合を抽出」の「部分集合の表現」で SEX=="F",「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**「統計量」→「要約」→「相関の検定」**を実行すればよい。

```
Pearson's product-moment correlation

data:  Females$HT and Females$WT
t = 7.1667, df = 47, p-value = 4.569e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5539837 0.8342857
sample estimates:
      cor
0.7226128
```

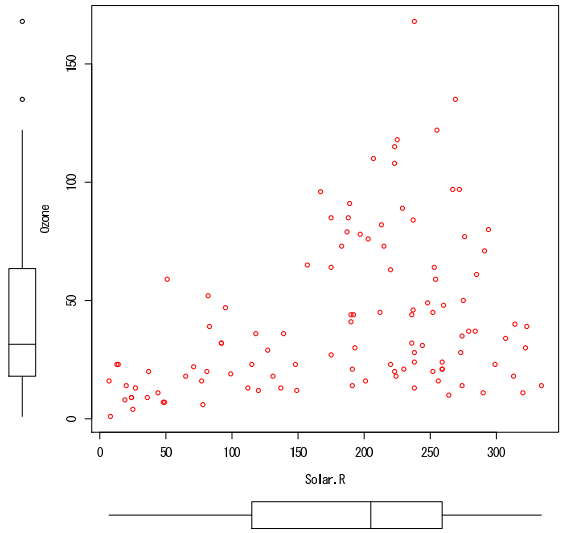
以上より，身長と体重の相関係数は，男性で 0.786（95% 信頼区間が [0.647,0.875]），女性で 0.723（95% 信頼区間が [0.554,0.834]）とわかった。男女とも統計学的に有意な強い正の相関があったといえる。

このデータでは必要なかったが，相関関係が直線的でなかったり，外れ値があったりする場合は，順位相関係数を使うのが適切な場合もある。

例題 3

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。太陽放射の強さとオゾン濃度の相関関係を検討せよ。

まずデータをアクティブにして散布図を描くのはいつも同じである。「データ」の「パッケージ内のデータ」の「アタッチされたパッケージからデータセットを読み込む」を選び,「データセット名を入力」のボックスに airquality と打って OK ボタンをクリックしてから,「グラフ」の「散布図」で「x 変数」として「Solar.R」を選び,「y 変数」として「Ozone」を選び,「最小 2 乗直線」と「平滑線」のチェックを外してから OK ボタンをクリックする。



どう見ても直線的な関係とは言いがたいので，スピアマンの順位相関係数を計算して，その有意性の検定をしてみる。「統計量」「要約」「相関の検定」で変数として Solar.R と Ozone を選び，相関のタイプを「スピアマンの順位」にして OK ボタンをクリックすると，次の結果が得られる。弱いけれども有意な相関があるといえる。

```
Spearman's rank correlation rho

data:  airquality$Ozone and airquality$Solar.R
S = 148561.3, p-value = 0.0001806
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3481865
```

もっとも，このデータの場合はピアソンの積率相関係数でも似たような結果が得られ（各自確かめよ），直線的な相関でないことの影響はあまりクリアでない。

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。

[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。

1.5 回帰モデルの数理

既に述べたとおり，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。いずれも，量がわかっているもの（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（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。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[*1]。

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。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，

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

が得られる[*2]。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。

[*1] 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。

[*2] 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。

1.6 回帰モデルの当てはまり

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，

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

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

として得られる Q は，回帰直線の当てはまりの悪さを示す尺度となる。Q を「残差平方和」と呼び，それを n で割った Q/n を残差分散という。この残差分散（$\mathrm{var}(e)$ と書くことにする）と Y の分散 $\mathrm{var}(Y)$ 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。

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。

以上の説明からすると，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，一方を独立変数，他方を従属変数とすることは，本当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，できれば避けたい。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。従って，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，ということは，因果関係の方向性に基づいて（先行研究や biological なメカニズムを参照して）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）。

例題 4

例題 3 のニューヨーク大気環境データについて，日照の強さを独立変数，オゾン濃度を従属変数とする回帰モデルを立てて分析せよ。

既に散布図は描いたが，回帰分析の場合は最小 2 乗直線も描くのが普通なので，そこをやり直す。次に，「統計量」の「モデルへの適合」の「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と，「出力ウィンドウ」に次の結果が得られる。

```
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 $\mathrm{Ozone} = 18.599 + 0.127 \cdot \mathrm{Solar.R}$ 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。

1.7 共分散分析

複数のグループがあって，どのグループに属するサンプルについても，同じ独立変数と従属変数が調べられているときに，独立変数と従属変数の関係がグループによって異なるかどうか調べたい場合がある。共分散分析は，このような場合に用いることができる分析手法である。

典型的には，$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$ というモデルになる。2 値変数 X_2 によって示される 2 群間で，量的変数 Y の平均値に差があるかどうかを比べるのだが，Y が量的変数 X_1 と相関がある場合に（このとき X_1 を共変量と呼ぶ），X_1 と Y の回帰直線の傾き (slope) が X_2 の示す 2 群間で差がないときに，X_1 による影響を調整した Y の修正平均（adjusted mean; 調整平均ともいう）に，X_2 の 2 群間で差があるかどうかを検定する。修正平均は X_2 の各変数についての係数（2 群の場合，基準にする変数の係数はゼロ）に，共変量の平均に共変量の係数を掛けたものを加え，さらに切片を加えることによって計算できる。

ただし，この検定をする前に，2 本の回帰直線がともに有意にデータに適合していて，かつ 2 本の回帰直線の間で傾き (slope) が等しいかどうかを検定して，傾きが等しいことを確かめておかないと，修正平均の比較には意味がない。そもそも回帰直線の適合が悪ければその独立変数は共変量として考慮する必要がないし，傾きが違っていれば群分け変

数と独立変数の交互作用が従属変数に関して有意に影響しているということなので，２群を層別して別々に解釈する方が良い。

いま，C で群分けされる２つの母集団における，(X, Y) の間の母回帰直線を，$y = \alpha_1 + \beta_1 x$，$y = \alpha_2 + \beta_2 x$ とすれば，次の２つの仮説が考えられる。まず傾きに差があるかどうか？　を考える。つまり，$H_0 : \beta_1 = \beta_2$，$H_1 : \beta_1 \neq \beta_2$ である。次に，もし傾きが等しかったら，y 切片も等しいかどうかを考える。つまり，$\beta_1 = \beta_2$ のもとで，$H_0' : \alpha_1 = \alpha_2$，$H_1' : \alpha_1 \neq \alpha_2$ を検定する。各群について，X と Y の平均と変動と共変動を出しておけば[*3]，仮説 H_1 のもとでの残差平方和

$$d_1 = SS_{Y1} - (SS_{XY1})^2/SS_{X1} + SS_{Y2} - (SS_{XY2})^2/SS_{X2}$$

と仮説 H_0 のもとでの残差平方和

$$d_2 = SS_{Y1} + SS_{Y2} - (SS_{XY1} + SS_{XY2})^2/(SS_{X1} + SS_{X2})$$

を計算して $F = (d_2 - d_1)/(d_1/(N-4))$ が H_0 のもとで第 1 自由度 1，第 2 自由度 $N-4$ の F 分布に従うことを使って傾きが等しいかどうかの検定ができる。H_0 が棄却されたときは，$\beta_1 = SS_{XY1}/SS_{X1}$，$\beta_2 = SS_{XY2}/SS_{X2}$ として別々に傾きを推定し，y 切片 α もそれぞれの式に各群の平均値を入れて計算できる。H_0 が採択されたときは，共通の傾き β を，$\beta = (SS_{XY1} + SS_{XY2})/(SS_{X1} + SS_{X2})$ として推定する。この場合はさらに y 切片が等しいという帰無仮説 H_0' のもとで全部のデータを使った残差平方和 $d_3 = SS_Y - (SS_{XY})^2/SS_X$ を計算して，$F = (d_3 - d_2)/(d_2/(N-3))$ が第 1 自由度 1，第 2 自由度 $N-3$ の F 分布に従うことを使って検定できる。H_0' が棄却された場合は各群の平均を共通の傾きに代入すれば各群の切片が求められるし，採択されたら，要するに２群間に差がないということになるので，２群を一緒にして普通の単回帰分析をしていいことになる。

例題 5

組み込みデータ `swiss`（1888 年頃のスイスのフランス語を話す 47 州についての，標準化された出生力水準 `Fertility`，農業就業割合 `Agriculture`，陸軍の試験で最高ランクを記録した人の割合 `Examination`，初等教育を超える教育を受けた人の割合 `Education`，カソリック信者割合 `Catholic`，乳児死亡割合 `Infant.Mortality` からなるデータ）を使って，教育水準が高いほど出生力が低いけれども，それがカソリック信者割合に影響を受ける（カソリック信者の方がプロテスタント信者よりも一般に出生力が高い）という仮説を検討してみよう。

「データ」→「パッケージ内のデータ」→「アタッチされたパッケージからデータセットを読み込む」として，パッケージとして `datasets`，データとして `swiss` を選択する。次に，**「データ」→「アクティブデータセット内の変数の管理」→「新しい変数を計算」**として，「新しい変数名」を `MoreCatholic`，計算式を `Catholic>=50` とすれば，カソリック信者割合が 50% 以上の州で TRUE，そうでない州で FALSE となる新しい論理型の変数 `MoreCatholic` ができる。しかし，Rcmdr では論理型の変数がカテゴリ変数とみなされないので，さらに「データ」→「アクティブデータセット内の変数の管理」→「変数の再コード化」で「再コード化の変数」として `MoreCatholic` を選び，「New variable name or prefix for multiple recodes」に `CatholicShare`，計算式を `FALSE="Less";TRUE="More"`として OK すると，やっと層別に使える因子型の変数 `CatholicShare` が得られる。

（注）このプロセスは，R Console では 1 行でできる。次のどちらかで良い（上の場合は水準名が”Less”と”More”ではなく，”[0,50)”と”[50,101)”になるが）。

```
swiss$CatholicShare <- cut(swiss$Catholic,c(0,50,101),right=F)
swiss$CatholicShare <- as.factor(ifelse(swiss$Catholic>=50,"More","Less"))
```

次に `CatholicShare` で層別して散布図を描かせ，最小 2 乗直線も層別に描かせる。つまり，「グラフ」→「散布図」から，「x 変数」として `Education`，「y 変数」として `Fertility` を選び，平滑線のチェックを外し，「層別のプロット」をクリックして `CatholicShare` を選んでから元のウィンドウで OK すれば目的のグラフが得られる。グラフを見ると最小 2 乗直線が交差しているので交互作用がありそうにも見える。

[*3] サンプルサイズ $N1$ の第 1 群に属する x_i, y_i について，$E_{X1} = \sum x_i/N1$，$SS_{X1} = \sum (x_i - E_{X1})^2$，$E_{Y1} = \sum y_i/N1$，$SS_{Y1} = \sum (y_i - E_{Y1})^2$，$E_{XY1} = \sum x_i y_i/N1$，$SS_{XY1} = \sum (x_i y_i - E_{XY1})^2$。第 2 群も同様。

次に層別に回帰分析をして線型回帰モデルが有意に当てはまるか調べる。**「統計量」→「モデルへの適合」→「線形回帰」**で「目的変数」として Fertility,「説明変数」として Education を選び,「部分集合の表現」として CatholicShare=="Less"として OK したときと CatholicShare=="More"として OK したときの出力を両方見ると,どちらでも Education の係数は 5% 水準で有意であることがわかる。

そこで,「統計量」→「モデルへの適合」→「線型モデル」で,モデル(モデル名 LinearModel.1)として左辺に Fertility を,右辺に Education*CatholicShare を指定すれば交互作用項により傾きの差を検討することができる。この場合,交互作用項が 5% 水準で有意なので,カソリック信者の多い郡と少ない郡の間では,教育水準と出生力の関係が異なっているといえ,層別回帰の結果を採用することになる。

ちなみに,もし傾きの差が有意でなければ,もう一度「線型モデル」を呼び出して,モデル名 LinearModel.2 とし,右辺を Education+CatholicShare として,CatholicShare の係数がゼロと有意差があるかどうかをみれば,カソリックが多い郡と少ない郡の間で教育水準の影響を調整した出生力の修正平均に差があるかどうかがわかる。修正平均そのものは Rcmdr では得られないので,R Console で以下を打つ。

```
cfs <- dummy.coef(LinearModel.2)
cfs[[1]] + cfs$Education * mean(swiss$Education) + cfs$CatholicShare
```

例題 6

http://phi.med.gunma-u.ac.jp/grad/sample3.dat は,都道府県別のタブ区切りテキストデータファイルである。変数としては,都道府県名 (PREF),日本の東西(REGION),1990 年の 100 世帯当たり乗用車台数(CAR1990),1989 年の人口 10 万人当たり交通事故死者数 (TA1989),1985 年の国勢調査による人口集中地区居住割合 (DIDP1985) が含まれている(REGION の 1 は東日本,2 は西日本を意味する)。
このデータについて,東日本と西日本で,100 世帯当たり乗用車台数で調整した人口 10 万人当たり交通事故死者数に差があるか,共分散分析によって検討せよ[a]。

[a] (注)実は乗用車台数の影響を調整しなければ人口当たり交通事故死者数は東西で有意な差はない。

データセット名 sample3 として web からデータを読み込む。まず REGION で層別した散布図を描く。「x 変数」を CAR1990,「y 変数」を TA1989 とすると,2 本の回帰直線はほぼ平行に見える。次に東西日本別々に層別して,CAR1990 によって TA1989 が説明されるかをみるため,単回帰分析を行う(やり方は例題 5 と同じなので省略する)。CAR1990 の係数は東西どちらでも有意にゼロと異なる。したがって,その影響を調整することに意味はあると思われる。

そこで,次に,傾きに差があるかを解析する。「統計量」→「モデルへの適合」→「線型モデル」でモデル名 LinearModel.3 として左辺に TA1989 を,右辺に CAR1990*REGION を指定する。結果をみると,交互作用効果は有意でないので,2 本の回帰直線の傾きに有意差はないことがわかる。

そこで今度は,乗用車所有台数で調整した交通事故死者数の修正平均に差があるかどうかをみるため,交互作用項を除いて回帰を行う。モデル名を LinearModel.4 とし,右辺を CAR1990+REGION として線型モデルの当てはめを実行する。この結果,REGION の効果は有意なので,乗用車保有台数で調整すると西日本の方が東日本よりも人口当たり交通事故死者数が多いことがわかる。修正平均は以下の枠内を R Console に打てば得られる。単純な平均値は東日本が 10.5,西日本が 9.87 であるが,乗用車保有台数の影響を調整した修正平均は,東日本が 9.44,西日本が 11.0 と逆転し,かつ有意差があることがわかる。

```
cfs <- dummy.coef(LinearModel.4)
cfs[[1]] + cfs$CAR1990 * mean(sample3$CAR1990) + cfs$REGION
```

1.8 ロジスティック回帰分析

ロジスティック回帰分析は,従属変数(ロジスティック回帰分析では反応変数と呼ぶこともある)が 2 値変数であり,二項分布に従うので lm() ではなく,glm() を使う一般化線型モデルとなる。ロジスティック曲線とは関係ない。従属変数がポアソン分布に従う場合も glm() で扱えるが,それはポアソン回帰と呼ばれる。

ロジスティック回帰分析の思想としては，例えば疾病の有無を，複数のカテゴリ変数によって表される要因の有無で説明する（量的な変数によって表される交絡を調整しながらオッズ比を計算できるのが利点であり，医学統計ではもっともよく使われる手法の一つである）。

この問題は，疾病の有病割合を P とすると，$\ln(P/(1-P)) = b_0 + b_1X_1 + ...b_kX_k$ と定式化できる。X_1 が要因の有無を示す2値変数で，$X_2, ...X_k$ が交絡であるとき，$X_1 = 0$ の場合を $X_1 = 1$ の場合から引けば，

$$b1 = \ln(P_1/(1-P_1)) - \ln(P_0/(1-P_0)) = \ln(P_1 * (1-P_0)/(P_0 * (1-P_1)))$$

となるので，b_1 が他の変数の影響を調整したオッズ比の対数になる。対数オッズ比が正規分布するとすれば，オッズ比の 95% 信頼区間が

$$\exp(b_1 \pm 1.96 \times \mathrm{SE}(b_1))$$

として得られる。

例題 7

`library(MASS)` の `data(birthwt)` は，Springfield の Baystate 医療センターの 189 の出生について，低体重出生とそのリスク因子の関連を調べるためのデータであり，次の変数を含んでいる。

low	低体重出生の有無を示す2値変数（児の出生時体重 2.5 kg 未満が 1）
age	年齢
lwt	最終月経時体重（ポンド[a]）
race	人種（1＝白人，2＝黒人，3＝その他）
smoke	喫煙の有無（1＝あり）
ptl	非熟練労働経験数
ht	高血圧の既往（1＝あり）
ui	子宮神経過敏の有無（1＝あり）
ftv	妊娠の最初の3ヶ月の受診回数
bwt	児の出生時体重 (g)

[a] 略号 lb. で，1 lb. は 0.454 kg に当たる。

低体重出生の有無を反応変数としたロジスティック回帰分析をせよ。

データには多くの変数が含まれているが，本来，ロジスティック回帰分析では，従属変数に対する効果を見たい変数と交絡因子となっている変数はすべて独立変数としてモデルに投入するべきである。独立変数と従属変数の両方と有意な相関があれば交絡因子となっている可能性がある。独立変数が多いときはステップワイズ法（`step()` という関数がある）を使いたくなるかもしれないが，1 つずつ丁寧に吟味して決定するのが筋である。

ここでは，丁寧な考察を経て，独立変数が人種，喫煙の有無，高血圧既往の有無，子宮神経過敏の有無，最終月経時体重，非熟練労働経験数となったとしよう。ロジスティック回帰分析の前に，`birthwt` では，人種なども数値型なので，**要因型に変換**しておく。**「データ」「アクティブデータセット内の変数の管理」「数値変数を因子に変換」**を選び，まず変数として `low` を選び，そのまま（同じ変数名だと上書きするかどうか尋ねるダイアログが出てくるが無視してよい。ただし変換に失敗すると元の変数の内容も壊れることがある）OK ボタンをクリックする。数値 0 が水準 1 となり NBW と名付け，数値 1 が水準 2 となり，LBW と名付ける。次に `race` を選び，そのまま OK ボタンをクリックし，水準ごとにカテゴリ名をつけるウィンドウに対し，第 1 水準に”white”，第 2 水準に”black”，第 3 水準に”others”と指定し，OK ボタンをクリックする。`smoke`，`ht`，`ui` についても同様にカテゴリ変数にする。

ロジスティック回帰分析は，**「統計量」「モデルへの適合」「一般化線型モデル」**で，式の左辺に `low`（**因子**）をクリックして代入し（たんに `low` と入る），右辺に `race+smoke+ht+ui+lwt+ptl` と打つ（またはクリックして選ぶ）。リンク関数族を `binomial` にして，リンク関数を `logit` にして OK すると，出力ウィンドウに次の枠内が表示される。

```
Call:
glm(formula = low ~ race + smoke + ht + ui + lwt + ptl, family = binomial(logit),
    data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9049  -0.8124  -0.5241   0.9483   2.1812

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.086550   0.951760  -0.091  0.92754
race[T.black]        1.325719   0.522243   2.539  0.01113 *
race[T.other]        0.897078   0.433881   2.068  0.03868 *
smoke[T.smoke]       0.938727   0.398717   2.354  0.01855 *
ht[T.hypertension]   1.855042   0.695118   2.669  0.00762 **
ui[T.uterine-hyper]  0.785698   0.456441   1.721  0.08519 .
lwt                 -0.015905   0.006855  -2.320  0.02033 *
ptl                  0.503215   0.341231   1.475  0.14029
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 201.99  on 181  degrees of freedom
AIC: 217.99

Number of Fisher Scoring iterations: 4
```

この出力からロジスティック回帰分析の結果を表としてまとめるのだが，切片と量的な変数のオッズ比には意味が薄いので，表の中でなく下に共変量として調整したと書くのが普通である。また，このままでは係数が対数オッズのままなので，指数をとる必要がある。95% 信頼区間も表示するのが普通である。そこは残念ながら Rcmdr ではまだできないので，分析結果が付値されている変数（回帰分析のモデルを指定するウィンドウの中で「モデル名」として指定したもの）が `GLM.1` だったとすると，R コンソールで，`exp(coef(GLM.1))` とすればオッズ比の点推定量が，`exp(confint(GLM.1))` とすれば 95% 信頼区間が得られる。p 値は出力ウィンドウに表示されている。

表. Baystate 医療センターにおける低体重出生リスクのロジスティック回帰分析結果

		95% 信頼区間		
独立変数 *	オッズ比	下限	上限	p 値
人種（白人）				
黒人	3.765	1.355	10.68	0.011
他の有色人種	2.452	1.062	5.878	0.039
喫煙あり（なし）	2.557	1.185	5.710	0.019
高血圧既往あり（なし）	6.392	1.693	27.3	0.008
子宮神経過敏あり（なし）	2.194	0.888	5.388	0.085

Nagelkerke の R^2: 0.223，AIC: 217.99，D_{null}: 234.67（自由度 188），D: 201.99（自由度 181）

* カッコ内はリファレンスカテゴリ。これらの変数の他，最終月経時体重と非熟練労働経験数を共変量としてロジスティック回帰モデルに含んでいる。

例題 8

前回 2 群の平均値の差の検定で用いたデータセット infert を用いて，不妊かどうか (case) を従属変数，年齢 (age)，既往出生児数 (parity)，教育歴 (education)，人工妊娠中絶経験数 (induced)，自然流産経験数 (spontaneous) を独立変数とするロジスティック回帰分析を実行せよ。

Rcmdr で**「データ」→「パッケージ内のデータ」→「アタッチされたパッケージからデータセットを読み込む」**として，パッケージとして datasets，データとして infert を選択するところまでは前回と同じである。

次いで，**「統計量」→「モデルへの適合」→「一般化線型モデル」**として，出てくるウィンドウ上で，「モデル名を入力」は GLM.2 のまま（別の名前をつけても構わない），モデル式として左辺に case，右辺を age + parity + education + induced + spontaneous とし，リンク関数族（ダブルクリックで選択）のところを binomial，リンク関数を logit として OK ボタンをクリックすれば，下記の結果が得られる。

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7603  -0.8162  -0.4956   0.8349   2.6536

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.14924    1.41220  -0.814   0.4158
age                   0.03958    0.03120   1.269   0.2046
parity               -0.82828    0.19649  -4.215 2.49e-05 ***
education[T.6-11yrs] -1.04424    0.79255  -1.318   0.1876
education[T.12+ yrs] -1.40321    0.83416  -1.682   0.0925 .
induced               1.28876    0.30146   4.275 1.91e-05 ***
spontaneous           2.04591    0.31016   6.596 4.21e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 257.80  on 241  degrees of freedom
AIC: 271.8

Number of Fisher Scoring iterations: 4
```

この結果から，既往出生児数が多いほど不妊になりにくく，人工妊娠中絶や自然流産が多いほど不妊になりやすいことがわかる。R Console で exp(coef(GLM.2)) とすると，既往出生児数が 1 増えると不妊リスクは 0.44 倍になり，人工妊娠中絶経験数が 1 増えると不妊リスクが 3.6 倍，自然流産が 1 増えると不妊リスクが 7.7 倍になることがわかる。なお，ペアごとに不妊リスクが異なることを仮定した条件付ロジスティック回帰分析をするには survival ライブラリの clogit() 関数を用いる必要があるが，Rcmdr ではできないので説明は省略する。使い方は example(infert) を実行すればわかる。

2　2 つのカテゴリ変数間の関係

2 つの量的変数間の関係を調べるには相関をみればよいわけだが，カテゴリ変数間の関係を調べるにはどうしたらよいだろうか？　もちろん，カテゴリ変数についても関連の強さをみる指標はあって，ファイ係数（記号は ρ を用いるのが普通）と呼ばれる指標は，要因の有無，発症の有無を 1,0 で表した場合のピアソンの積率相関係数と同じ計算式で得られる。θ_1, θ_2 を発症者中の要因あり割合，非発症者中の要因あり割合として，$\rho = \sqrt{(\pi_1 - \pi_2)(\theta_1 - \theta_2)}$ である。また，疫学研究では，人数あるいは人年の比を取ることによって，要因があった群が，要因がなかった群に比べて，どれくらい発症しやすいかを調べることが多い（オッズ比やリスク比やハザード比を求め，その 95% 信頼区間が 1 を含ま

ないかどうかで，要因の有無が発症の有無に有意に影響しているかどうかを判定することが慣例的に行われる)。

しかし，2つのカテゴリ変数の関係を考えるとき，一般に，もっともよく行われるのは，それらが独立であるという帰無仮説を立てて検定することである。

カテゴリ変数のもつ統計的な情報は，カテゴリごとの度数だけである。そこで，2つのカテゴリ変数の間に関係について検討したいときには，まず**それらの組み合わせの度数**を調べた表を作成する（R では `table()` という関数が使える)。これをクロス集計表と呼ぶ。とくに，2つのカテゴリ変数が，ともに2値変数のとき，そのクロス集計は2×2クロス集計表（2×2分割表）と呼ばれ，その統計的性質が良く調べられている。

2.1 独立性のカイ二乗検定

独立性の検定としては，2つのカテゴリ変数の間に**関連がない**と仮定した場合に推定される期待度数を求めて，それに観測度数が適合するかを検定するカイ二乗検定が最も有名である（だから，実はカイ二乗適合度検定と同じ原理である)。もちろん，ある種の関連が仮定できれば，その仮定の元に推定される期待度数と観測度数との適合を調べてもいいが，一般に，2つのカテゴリ変数の間にどれくらいの関連がありそうかという仮定はできないことが多い。そこで，関連がない場合の期待度数を推定し，それが観測値に適合しなければ関連がないとはいえない，と推論するのである。

	A	$\bar{A}$
B	a 人	b 人
$\bar{B}$	c 人	d 人

2つのカテゴリ変数 A と B が，それぞれ「あり」「なし」の2つのカテゴリ値しかとらないとき，これら2つのカテゴリ変数の組み合わせは「A も B もあり（$A \cap B$)」「A なし B あり（$\bar{A} \cap B$)」「A あり B なし（$A \cap \bar{B}$)」「A も B もなし（$\bar{A} \cap \bar{B}$)」の4通りしかない。それぞれの度数を数えあげた結果が，上記の表として得られたときに，母集団の確率構造が，

	A	$\bar{A}$
B	π_{11}	π_{12}
$\bar{B}$	π_{21}	π_{22}

であるとわかっていれば，$N = a + b + c + d$ として，期待される度数は，

	A	$\bar{A}$
B	$N\pi_{11}$	$N\pi_{12}$
$\bar{B}$	$N\pi_{21}$	$N\pi_{22}$

であるから，

$$\chi^2 = \frac{(a - N\pi_{11})^2}{N\pi_{11}} + \frac{(b - N\pi_{12})^2}{N\pi_{12}} + \frac{(c - N\pi_{21})^2}{N\pi_{21}} + \frac{(d - N\pi_{22})^2}{N\pi_{22}}$$

として, 自由度3のカイ二乗検定をすればよいことになる。しかし, 普通, π は未知である。そこで, $Pr(\bar{A}) = 1 - Pr(A)$ かつ，この2つのカテゴリ変数が独立ならば $Pr(A \cap B) = Pr(A)Pr(B)$ と考えれば良い[*4]ことを使って，$Pr(A)$ と $Pr(B)$ を母数として推定する[*5]。$Pr(A)$ の点推定量は，B を無視して A の割合と考えれば $(a + c)/N$ であることは自明である。同様に，$Pr(B)$ の点推定量は，$(a + b)/N$ となる。したがって，$\pi_{11} = Pr(A \cap B) = Pr(A)Pr(B) = (a + c)(a + b)/(N^2)$ となる。

同様に考えれば，母集団の各セルの確率は下式で得られる。

$$\pi_{12} = (b + d)(a + b)/(N^2)$$

$$\pi_{21} = (a + c)(c + d)/(N^2)$$

[*4] この帰無仮説は，合計に比例する割合で人数配分が行われていることに相当するので，前回説明したとおり，B あり群と B なし群のそれぞれについて，A ありの割合に差がないという，比率の差の検定の帰無仮説と数学的に等価である。

[*5] $Pr(X)$ はカテゴリ X の出現確率を示す記号である。また，2つの母数をデータから推定するので，得られるカイ二乗統計量が従う分布の自由度は3より2少なくなり，自由度1のカイ二乗分布となる。

$$\pi_{22} = (b+d)(c+d)/(N^2)$$

これらの値を使えば,

$$\chi^2 = \frac{\{a-(a+c)(a+b)/N\}^2}{\{(a+c)(a+b)/N\}} + \frac{\{b-(b+d)(a+b)/N\}^2}{\{(b+d)(a+b)/N\}} + \frac{\{c-(a+c)(c+d)/N\}^2}{\{(a+c)(c+d)/N\}} + \frac{\{d-(b+d)(c+d)/N\}^2}{\{(b+d)(c+d)/N\}}$$

$$= \frac{(ad-bc)^2\{(b+d)(c+d)+(a+c)(c+d)+(b+d)(a+b)+(a+c)(a+b)\}}{(a+c)(b+d)(a+b)(c+d)N}$$

分子の中括弧の中は N^2 なので，結局,

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

ただし通常は，イェーツの連続性の補正を行う。カイ二乗分布は連続分布なので，各度数に 0.5 を足したり引いたりしてやると，より近似が良くなるという発想である。この場合,

$${\chi_c}^2 = \frac{N(|ad-bc|-N/2)^2}{(a+c)(b+d)(a+b)(c+d)}$$

が自由度 1 のカイ二乗分布に従うと考えて検定する。ただし，$|ad-bc|$ が $N/2$ より小さいときは補正の意味がないので，$\chi^2 = 0$ とする。

実際の検定はクロス集計表が既に得られているとき，例えば a=12, b=8, c=9, d=10 などとわかっていれば，**「統計量」「分割表」「２元表の入力と分析」で表示される表の各セルに直接数字を入力し，必要な統計量のチェックボックスにチェックを入れて OK ボタンをクリックするだけ**で非常に簡単である（注：前回，比率の差の検定のところで説明したとおり)。

例題 9

肺ガンの患者 100 人に対して，１人ずつ性・年齢が同じ健康な人を対照として 100 人選び（この操作をペアマッチサンプリングという)，それぞれについて過去の喫煙の有無を尋ねた結果，患者群では過去に喫煙を経験した人が 80 人，対照群では過去に喫煙を経験した人が 55 人だった。肺ガンと喫煙は無関係といえるか？　独立性のカイ二乗検定をせよ。

帰無仮説は，肺ガンと喫煙が無関係（独立）ということである。クロス集計表を作ってみると,

	肺ガン患者群	健康な対照群	合計
過去の喫煙経験あり	80	55	135
過去の喫煙経験なし	20	45	65
合計	100	100	200

となる。肺ガンと喫煙が無関係だという帰無仮説の下で期待される各カテゴリの人数は,

	肺ガンあり	肺ガンなし
喫煙あり	$135\times100/200 = 67.5$	$135\times100/200 = 67.5$
喫煙なし	$65\times100/200 = 32.5$	$65\times100/200 = 32.5$

となる。従って，連続性の補正を行なったカイ二乗統計量は,

$$\chi_c^2 = (80-68)^2/67.5 + (55-67)^2/67.5 + (20-32)^2/32.5 + (45-33)^2/32.5 = 13.128...$$

となり，自由度 1 のカイ二乗分布で検定すると `1-pchisq(13.128,1)` より有意確率は 0.00029... となり，有意水準 5% で帰無仮説は棄却される。つまり，肺ガンの有無と過去の喫煙の有無には有意な関連があるといえる。**生データがカテゴリ変数としてある場合は，「統計量」「分割表」「２元表」で２つのカテゴリ変数を選べばよい。**例題 7 のデータで `low` と `smoke` の独立性の検定をして試してみよ。

2.2 フィッシャーの直接確率（正確な確率）

期待度数が低い組み合わせがあるときには，カテゴリを併合して変数を作り直す方法もあるが，もっといい方法が考案されている。

ここで調べたいのは組み合わせの数なので，周辺度数を固定して（各々の変数については母比率が決まっていると仮定して）すべての組み合わせを考え，それらが起こる確率（超幾何分布に従う）を１つずつ計算し，得られている集計表が得られる確率よりも低い確率になるような表が得られる確率をすべて足し合わせてしまえば，２つのカテゴリ変数の間に関連がないという帰無仮説の下でそういう表が偶然得られる確率がどれほど低いのかを，直接計算することができる。こうして計算される確率を，フィッシャーの直接確率，あるいは，フィッシャーの正確な確率（検定）という。これなら，近似ではないので，期待度数が低い組み合わせがあっても問題ない。

もう少し丁寧に言うと，サイズ N の有限母集団があって，そのうち変数 A の値が 1 である個体数が m_1，1 でない個体数が m_2 あるときに，変数 B の値が 1 である個体数が n_1 個（1 でない個体数が $n_2 = N - n_1$ 個）あるという状況を考え，この n_1 個のうち変数 A の値が 1 である個体数がちょうど a である確率を求めることになる。これは，m_1 個から a 個を取り出す組み合わせの数と m_2 個から $n_1 - a$ 個を取り出す組み合わせの数を掛けて，N 個から n_1 個を取り出す組み合わせの数で割った値になる。これと同じ周辺度数をもつ２×２分割表のうち，確率がこれと同じかこれよりも小さい表の確率をすべて足し合わせたものが，「変数 A と変数 B が独立」という帰無仮説が成り立つ確率になる[*6]。

クロス集計表を使って２つのカテゴリ変数間の独立性の検定をするときは，コンピュータが使えるならば，サンプルサイズがよほど大きくない限り，常にカイ二乗検定ではなく，フィッシャーの正確な確率を求めるべきである。**Rcmdr では，２元表の分析で「フィッシャーの正確検定」にチェックを入れておけば計算してくれる。**

3 生存時間解析

生存時間解析は，まだ Rcmdr ではできないが，重要なのでごく簡単に説明を書いておく。

3.1 生存時間解析とは

実験においては，化学物質などへの１回の曝露の影響を時間を追ってみていくことが良く行われる。時間ごとに何らかの量の変化を追うほかに，エンドポイントを死亡とした場合，死ぬまでの時間を分析することで毒性の強さを評価することができる。このような期間データを扱う方法としては，一般に生存時間解析 (Survival Analysis または Event History Analysis) と呼ばれるものがある。なかでもよく知られているものが Kaplan-Meier の積・極限推定量である（現在では一般に，カプラン・マイヤ推定量と呼ばれている）。カプラン・マイヤ推定量は，イベントが起こった各時点での，イベントが起こる可能性がある人口（リスク集合）あたりのイベント発生数を１から引いたものを掛け合わせて得られる，ノンパラメトリックな最尤推定量である。カプラン・マイヤ推定量は非常によく使われるので，具体例で説明しておく（後述）。複数の期間データ列があったときに，それらの差を検定したい場合は，ログランク検定や一般化ウィルコクソン検定が使われる。細かいことをいえば，ログランク検定でも Mantel-Haenzel 流のログランク検定と Peto and Peto 流のログランク検定があったり，一般化ウィルコクソン検定でも Gehan-Breslow 流と Peto-Prentice 流があったりして，非常に面倒な話になってくるので，ここでは Mantel-Haenzel 流のログランク検定のみ[*7]説明する（後述）。

それらのノンパラメトリックな方法とは別に，イベントが起こるまでの時間が何らかのパラメトリックな分布に当て

[*6] 有限母集団からの非復元抽出になるので，平均 $E(a)$ と分散 $V(a)$ は，

$$E(a) = n_1 m_1 / N$$

$$V(a) = \{(N - n_1)/(N - 1)\} n_1 (m_1/N)(m_2/N) = (m_1 m_2 n_1 n_2)/\{N^2 (N - 1)\}$$

となる。実際には組み合わせ計算が多いので，手計算で実行することはまずありえず，統計ソフトにやらせることになる。また，個々の２×２分割表の確率は離散値をとるので，同じ確率の表がありうる場合に，それを足し算に含めるのかどうかは難しい点である。これを乱数によって決める「ランダム検定」という手法もあるが，あまり一般的ではない。

[*7] Peto and Peto 流のものは，コンピュータが使えない場合に手計算するには有用だが，それ以上の意味はない。

MO_ID	MO_BD	C1_BD	C2_BD	C3_BD	C4_BD	C5_BD	C6_BD	C7_BD	C8_BD	C9_BD	C10_BD	C11_BD
20102	390000	0	640600	680000	711014	760000						
60202	250000	480415	560921	630000								
50102	400000	530000	590000	630000	660810	681011	710319	741018	760611	0		
30602	450000	580000	601004	630000	650000	670000	670000	720000	740000	750000	780714	
10502	400000	600716	630000	650807	670000	690609						
10102	400000	651103	681225	0	720200	0	790517	0	820000	840503	860527	890302
30102	490000	680000	700000	720000	750000	770000	820927					
10202	490000	680000	720826	760000	830000							
40302	580000	700000	780606	820906	901012	910606						
40102	570000	710114	730000	750000	770000	810621	840101	870802	920813			
20502	580000	720906	740704	761106	800407	811126	860516	910406				
50302	520000	730000	780000	800000	830000	870000	0					
10402	441101	730324	760725	770801	880119							
60302	460000	740000	770000	790000	800000	820000						
70202	550000	740000	780000	800000	840000	870000	890000	920000	941100			
70302	600000	750000	780000	800000	820000	850500	860000	880000	920000	940000		
20302	610000	760709	771020	790309	811002	850415	890803					
30702	600000	810530	820000	830000	840000	850000	900924	930430	950604			
30502	501205	820921	840803	881228								
60402	530000	830212	850216	900916	930921							
10802	650521	840623	861009	890727	920329	940416						
50402	670000	861114	880430	900130	910000	930325	950108					
20602	651114	870904	881111	900519	911104							
60102	570000	880700	950905									
10902	670000	900000	910000	950319								
30202	710000	900408	920210	940305								
40202	640000	901007	931109									
60204	680000	910000	920000									
50202	640000	911001	921020									
20202	711014	920801	931127									
10302	720826	920823	940308									
11002	700917	930303	950513									
10702	670304	930701										
80102	720229	940125										
11102	670809	940406										
30302	720000	940611										
50303	730000	950300										
10602	740700	950317										
20504	740704	950905										
60303	740000	951024										
70102	0	420000	450000	470000	520000	531225	550000	630000	670000			

図 1　ソロモン諸島のある村の女性全員の再生産史

はまるかどうかを調べる方法もある。当てはめる分布としては指数分布やワイブル分布がある。イベントが起こるまでの期間に何らかの別の要因が与える効果を調べたいときはコックス回帰（それらが基準となる個体のハザードに対して $\exp(\sum \beta_i z_i)$ という比例定数の形で掛かるとする比例ハザード性を仮定する方法）と，パラメトリックなモデルに対数線形モデルの独立変数項として入れてしまう加速モデルがある。R では生存時間解析をするための関数は `survival` パッケージで提供されており，`library(survival)` とすれば使えるようになる。カプラン・マイヤ法は `survfit()` 関数，コックス回帰は `coxph()` 関数，加速モデルは `survreg()` 関数で実行できる。なお，生存時間解析について，より詳しく知りたい方は，大橋，浜田 (1995) などを参照されたい。

3.2　カプラン・マイヤ法

では，簡単な例を使って，カプラン・マイヤ推定量を説明しよう。表 1 は，ソロモン諸島のある村で，既婚女性全員に，自分の誕生日，第 1 子誕生日，第 2 子誕生日，……，末子誕生日（まだ出産を完了していない年齢の女性も含めて，ともかくそれまでに産んだ子どもの誕生日を全部）聞き取った結果である。間隔データを使わなければ，このデータから出生力について何かいうためには，出産を完了した女性についての平均出産数（平均完結パリティという）くらいしか指標がないが，間隔データを使えば，時間当たりの出生力を考えることができるので，出産を完了していない女性のデータも使うことができる。

この種のデータには，以下の利点と欠点がある。

- 母親に対して，全ての子どもの出生年月日を聞き取ることは，統計がしっかりしていない社会でも比較的信頼性の高い方法である。
- 人口規模が小さくても使える上，過去の推計もできるという利点がある。
- 古くなるほど誤差が大きくなるバイアスや，他に影響を受ける要因が多いのは欠点。
- 結婚から第 1 子誕生までの期間や，第 1 子と第 2 子の出生間隔がよく使われるが，上にあげたソロモン諸島の社会では，結婚記念日はあまり正確に記憶されていなかったために，第 1 子と第 2 子の出産間隔を使うことにした。第 1 子と第 2 子の出産間隔には，第 2 子の在胎期間が含まれるために，その期間のハザードは原理的にゼロであることに注意する必要がある[*8]。

まず，カプラン・マイヤ推定量についての一般論を示す。イベントが起こる可能性がある状態になってから，イベントが起こった時点を $t_1, t_2, ...$ とし，t_1 時点でのイベント発生数を d_1，t_2 時点でのイベント発生数を d_2，以下同様であ

[*8] 例えば，在胎期間の推定値として 9 ヶ月を引いた値をデータにしたり，または在胎期間を切片として含んだハザード関数を推定することも考慮するべきである。

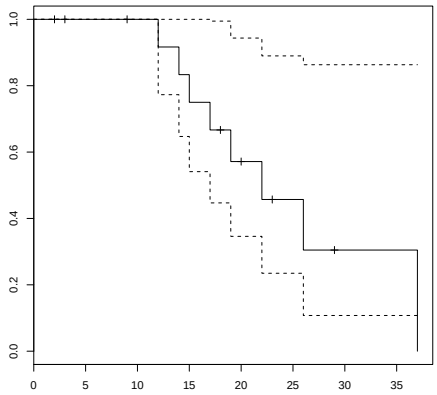


図 2　ソロモン諸島女性の第 1 出産間隔についてのカプラン・マイヤプロット

るとする。また，時点 $t_1, t_2, ...$ の直前でのリスク集合の大きさを $n_1, n_2, ...$ で示す。リスク集合の大きさとは，その直前でまだイベントが起きていない（この例では第 1 子出産後で第 2 子出産前の）個体数である。観察途中で死亡や転居などによって打ち切りが生じるために，リスク集合の大きさはイベント発生によってだけではなく，打ち切りによっても減少する。従って n_i は，時点 t_i より前にイベント発生または打ち切りを起こした個体数を n_1 から除いた残りの数となる。なお，イベント発生と打ち切りが同時点で起きている場合は，打ち切りをイベント発生直後に起きたと見なして処理するのが慣例である。このとき，カプラン・マイヤ推定量 $\hat{S}(t)$ は，

$$\hat{S}(t) = (1 - d_1/n_1)(1 - d_2/n_2)... = \prod_{i<t}(1 - d_i/n_i)$$

として得られる。その標準誤差はグリーンウッドの公式により（説明は省略するが），

$$var(\hat{S}) = \hat{S}^2 \times \sum_{i<t} \frac{d_i}{n_i(n_i - d_i)}$$

で得られる。なお，カプラン・マイヤ推定量を計算するときは，階段状のプロットを同時に行うのが普通である。

R では，`library(survival)` としてパッケージを呼び出し，`dat <- Surv(`**生存時間，打ち切りフラグ**`)` 関数で生存時間データを作り（打ち切りフラグは 1 でイベント発生，0 が打ち切り。ただし区間打ち切りの場合は 2 とか 3 も使う），`res <- survfit(dat)` でカプラン・マイヤ法によるメディアン生存時間が得られ，`plot(res)` とすれば階段関数が描かれる。イベント発生時点ごとの値を見るには，`summary(res)` とすればよい[*9]。

例えば，区間打ち切り（イベント発生までの時間がある幅をもってしかわからないデータ）を無視して，上で示したソロモン諸島のデータのうち，第 1 子出生が 1986 年以降のものの出産間隔データを R で分析すると，右側打ち切りを考慮した出産間隔のメディアンが 22 ヶ月であることがわかる（プロットを図 2 に示す）。プログラムは次の枠内。

[*9] 参考までに書いておくと，生データがイベント発生の日付を示している場合，間隔を計算するには `difftime()` 関数や `ISOdate()` 関数を使う。例えば 1964 年 8 月 21 日生まれの人の今日の年齢は `integer(difftime(ISOdate(2007,6,13),ISOdate(1964,8,21)))/365.24` とすれば得られるし，さらに 12 を掛ければ月単位になる。

```
library(survival)
time <- c(17,14,22,37,12,15,19,26,29,23,20,18,9,9,3,2)
event <- c(1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0)
dat <- Surv(time,event)
res <- survfit(dat)
print(res)
summary(res)
plot(res)
```

3.3 ログランク検定

次に，ログランク検定を簡単な例で説明する。

8 匹のラットを 4 匹ずつ 2 群に分け，第 1 群には毒物 A を投与し，第 2 群には毒物 B を投与して，生存期間を追跡したときに，第 1 群のラットが 4,6,8,9 日目に死亡し，第 2 群のラットが 5,7,12,14 日目に死亡したとする。この場合，観察期間内にすべてのラットが死亡し，正確な生存時間がわかっているので，観察打ち切りがないデータとなっていて計算しやすい。

ログランク検定の思想は，大雑把に言えば，死亡イベントが起こったすべての時点で，群と生存／死亡個体数の 2 × 2 クロス集計表を作り，それをコクラン＝マンテル＝ヘンツェル流のやり方で併合するということである。上記の例では，死亡イベントが起こった時点 1〜8 において各群の期待死亡数を計算し，各群の実際の死亡数との差をとって，それに時点の重みを掛けたものを，各時点における各群のスコアとして，群ごとのスコアの合計を求める。2 群しかないので，各時点において群 1 と群 2 のスコアの絶対値は同じで符号が反対になる。2 群の生存時間に差がないという帰無仮説を検定するためには，群 1 のスコアの 2 乗を分散で割った値をカイ二乗統計量とし，帰無仮説の下でこれが自由度 1 のカイ二乗分布に従うことを使って検定する。なお，重みについては，ログランク検定ではすべて 1 である。一般化ウィルコクソン検定では，重みを，2 群を合わせたリスク集合の大きさとする（そうした場合，もし打ち切りがなければ，検定結果は，ウィルコクソンの順位和検定の結果と一致する）。つまり，ログランク検定でも一般化ウィルコクソン検定でも，実は期間の情報はまったく使われず，死亡順位の情報だけが使われているのである。

記号で書けば次の通りである。第 i 時点の第 j 群の期待死亡数 e_{ij} は，時点 i における死亡数の合計を d_i，時点 i における j 群のリスク集合の大きさを n_{ij}，時点 i における全体のリスク集合の大きさを n_i とすると，

$$e_{ij} = d_i \cdot n_{ij}/n_i$$

と表される。上の例では，$e_{11} = 1 \cdot n_{11}/n_1 = 4/8 = 0.5$ となる。時点 i における第 j 群の死亡数を d_{ij}，時点の重みを w_i と表せば，時点 i における群 j のスコア u_{ij} は，

$$u_{ij} = w_i(d_{ij} - e_{ij})$$

となり，ログランク検定の場合（以下，重みは省略してログランク検定の場合のみ示す）の群 1 の合計スコアは

$$u_1 = \sum_i d_{i1} - e_{i1}$$

となる。上の例では，

$$u_1 = (1 - 4/8) + (0 - 3/7) + (1 - 3/6) + (0 - 2/5) + (1 - 2/4) + (1 - 1/3) + (0 - 0/2) + (0 - 0/1)$$

である。これを計算すると約 1.338 となる。分散は，分散共分散行列の対角成分を考えればいいので，

$$V = V_{jj} = \sum_i \frac{(n_i - n_{ij})n_{ij}d_i(n_i - d_i)}{{n_i}^2(n_i - 1)}$$

となる。この例の数値を当てはめると，

$$V = \frac{(8-4)\times 4}{8^2} + \frac{(7-3)\times 3}{7^2} + \frac{(6-3)\times 3}{6^2} + \frac{(5-2)\times 2}{5^2} + \frac{(4-2)\times 2}{4^2} + \frac{(3-1)\times 1}{3^2}$$

となり，計算すると，約 1.568 となる。したがって，$\chi^2 = 1.338^2/1.568 = 1.14$ となり，この値は自由度 1 のカイ二乗分布の 95% 点である 3.84 よりずっと小さいので，有意水準 5% で帰無仮説は棄却されない。つまりこれだけのデータでは，差があるとはいえないことになる（もちろん，サンプルサイズを大きくすれば違う結果になる可能性もある）。

R では，`Surv(time,event)` と `group`（注：ここで time は生存時間，event は 1 がイベント観察，0 が観察打ち切りを示すフラグ，group がグループを示す）を，`survdiff()` 関数に与えることによってログランク検定が実行できる。打ち切りレコードがない場合は，event は省略できる。なお，生存時間解析の関数は `survival` パッケージに入っているので，まず `library(survival)` とすることは必須である。

上で説明した例では，

```
library(survival)
time <- c(4,6,8,9,5,7,12,14)
event <- c(1,1,1,1,1,1,1,1)
group <- c(1,1,1,1,2,2,2,2)
dat <- Surv(time,event)
res <- survfit(dat~group)
print(res)
summary(res)
res2 <- survdiff(dat~group)
print(res2)
```

とすると解析結果が得られる。一番下を見ると，$\chi^2 = 1.2$，自由度 1，$p = 0.268$ となっているので，有意水準 5% で，2 群には差がないことがわかる。

4 レポート課題

`http://phi.med.gunma-u.ac.jp/grad/sample4.dat` は，web サイトで zip 形式に圧縮して公開されている「The world factbook 2007」(CIA)[*10]から作ったタブ区切りテキスト形式のデータである。

変数は，COUNTRY（国名），GINI（Gini の集中係数）,YEAR.x（Gini の集中係数の報告年），LIFEEXP（ゼロ歳平均余命），YEAR.y（ゼロ歳平均余命の報告年），GDPPCUSD（米ドル換算購買力平価ベース 1 人当たり GDP），YEAR（GDP 報告年），RP（GDP が 10000 以上だと RICH，10000 未満だと POOR となるカテゴリ変数），EU（GINI の係数が 35 以上だと Uneven（不平等），35 未満だと Egaritarian（平等主義者）となるカテゴリ変数）である。

近年，社会疫学という研究分野において，健康が社会のありようによって影響を受けることが指摘されており，ゼロ歳平均余命が所得の不平等度や国内総生産から受ける影響を調べることも行われているが，このデータをその視点で統計解析せよ。図示や記述統計をしてから，豊かな国と貧しい国の間で Gini の集中係数とゼロ歳平均余命の関係の違いを調べよ。違いがない場合には Gini の集中係数を共変量としてその影響を調整したゼロ歳平均余命の修正平均を，豊かな国と貧しい国で比較せよ。

提出期限は 6 月末日までとする。A4 のレポート用紙で 3 枚以内にまとめて，基礎棟 5 階の公衆衛生学・中澤准教授室前のボックスに提出されたい。

[*10] `https://www.cia.gov/library/publications/the-world-factbook/index.html`