医学情報処理演習第 12 回「一般化線形モデル」[*1]

2006 年 1 月 23 日 中澤　港 (nminato@med.gunma-u.ac.jp)

演習サポート web ページ：http://phi.med.gunma-u.ac.jp/medstat/

前回の資料の訂正箇所について

前回の配布資料にはミスがありましたので以下の通り訂正します（web 上のものは修正済み)。

- (p.5, 下から 4 行目の分散の式）は 2 ヶ所誤りがあり，正しくは,

$$\begin{aligned} V(R_Y) = &\quad 32 \times 32 \times (32+32+1)/12 \\ &- \quad 32 \times 32/\{12 \times (32+32-1) \times (32+32)\} \times \{17^3 - 17 + (2^3-2) \times 5\} \\ &= \quad 5442.413 \end{aligned}$$

- (p.5，次の式の右辺) は，2.459507 ではなく，2.460259 となります。
- (p.5，最下行）は，0.01391280 ではなく，0.01388366 となります。
- (p.5，脚注）は,「`wilcox.test(X,Y,exact=F)` の結果の p 値は 0.1388 と表示され……」とあり，余計なことが書いてありますが,「`wilcox.test(X,Y,exact=F)` の結果と一致する」で OK です。
- (p.6, 枠内の VRY の計算式) は,
  `VRY <- 32*32*N/12-32*32/(12*(N-2)*(N-1))*(17^3-17+(2^3-2)*5)`
  が正しいです（`N <- 32+32+1` としたのを式の途中で忘れていました)。

前回の課題の回答例

帰無仮説「**9:00 と 21:00 の血清鉄濃度に差がない**」を，ノンパラメトリックな方法で検定するには,「**Wilcoxon の符号付き順位和**検定」を実行すればよいので,

```
BX <- c(0.98,0.87,1.12,1.34,0.88,0.91,1.04,1.21,1.17,1.09)
AX <- c(1.03,0.78,1.04,1.52,0.97,0.84,1.32,1.12,1.09,1.32)
```

として，9:00 の血清鉄濃度を `BX` に，21:00 の血清鉄濃度を `AX` に付値してから,

```
wilcox.test(BX,AX,paired=T,exact=F)
```

として検定を実行する。（下枠内の結果が得られるので,）結果の p-value をみると，**0.5073 と 5%よりずっと大きい**ので，有意水準 5%で帰無仮説は「**採択**」され，9:00 と 21:00 の血清鉄濃度に有意差は「**ないといえる**」。

```
        Wilcoxon signed rank test with continuity correction

data:  BX and AX
V = 20.5, p-value = 0.5073
alternative hypothesis: true mu is not equal to 0
```

[*1] 本資料は http://phi.med.gunma-u.ac.jp/medstat/it12-2005.pdf としてダウンロード可能である。

一般化線型モデルとは？

前回はなるべく仮定なしにデータを分析する方法を説明したが，今回のテーマは，逆に，かなり強い法則性を仮定して立てたモデルを，データに当てはめることである。モデルによってデータのすべてが完全に説明されることはまずありえないが，かなりの程度説明されれば，そのモデルはデータに内在する法則性の妥当な解釈を与えると判断できる。具体的なモデルとしては，重回帰，共分散，ロジスティック回帰を扱う。一般化線型モデル (Generalized Linear Model) は，基本的には，

$$Y = \beta_0 + \beta X + \varepsilon$$

という形で表される（Y が従属変数群[*2]，X が独立変数群（及びそれらの交互作用項），β_0 が切片群，β が係数群，ε が誤差項である）。係数群は未定であり，そのモデルがもっとも良くデータに当てはまるようになる数値を，最小二乗法または最尤法で求めるのが普通である。こうして得られる係数は，通常，偏回帰係数と呼ばれ，互いに他の独立変数の影響を調整した，各独立変数独自の従属変数への影響を示す値と考えられる（なお，相対的にどの独立変数の影響が大きいかをみるときは，独立変数の絶対値に依存してしまう偏回帰係数で比較することはできず，標準化偏回帰係数を用いる[*3]）。R では，`glm()` という関数を使ってモデルを記述するのが基本だが，外部ライブラリとして，もっと凝ったモデル記述とその当てはめを行うためのパッケージがいくつも開発され，CRAN で公開されている。また，一般化線型モデルとは違うモデルとして，独立変数群の効果が線型結合でない（例えば，ある独立変数の二乗に比例した大きさの効果があるような場合），いわゆる非線型モデルも `nls()` という関数で扱うことができる。

モデルの記述法

R の `glm()` 関数における一般化線型モデルの記述は，例えば，(1) 独立変数群が `X1` と `X2` で，従属変数が `Y` であり，`Y` が正規分布に従う場合，(2)(1) と同じ構造だが切片がゼロとして係数を推定したい場合，(3)`dat` というデータフレームに従属変数 `Y` と，その他すべての変数が独立変数として含まれていて，`Y` が 2 値変数である場合，(4) 独立変数群がカテゴリ変数 `C1`，`C2` と，それらの交互作用項で，従属変数が正規分布に従う量的変数 `Y` である場合，について順に示すと，下枠内のようになる。

```
glm(Y ~ X1+X2)
glm(Y ~ X1+X2-1)
glm(Y ~ ., data=dat, family="binomial")
glm(Y ~ C1+C2+C1:C2)
```

`family` のデフォルトは`"gaussian"`なので，上 2 行のように `family` を指定しなければ正規分布を仮定することになる。この場合，モデルとしては単純な線型重回帰モデルとなるため，例え

[*2] 変換したものである場合もある

[*3] なお，標準化偏回帰係数は，各偏回帰係数に各独立変数の標準偏差を掛け，従属変数の標準偏差で割れば得られる。

ば (1) の場合なら `lm(Y ~ X1+X2)` と同等である。`summary(lm())` ならば自由度調整済み重相関係数の二乗が得られるので，従属変数にも正規母集団を仮定できる，単純な線型重回帰で済むときは，`lm()` を使うことを薦める。(4) も従属変数が正規分布に従うので，`lm()` の方がよい。また，独立変数が複数のカテゴリ変数であるときに，主効果と交互作用項のすべてを指定するには，`*`で変数名をつなぐ方法もあり，(4) の右辺は `C1*C2` とも書ける。(4) のモデルは二元配置分散分析なので，結局，`anova(lm(Y ~ C1*C2))` とするのが普通である（脚注も参照せよ）。

また，これらのモデルの当てはめの結果は，`res <- glm(Y ~ X1+X2)` のようにオブジェクトに保存しておくことができ，`plot(residuals(res))` として残差プロットをしたり，`summary(res)` として詳細な結果を出力させたり，`AIC(res)` として AIC を計算させたり，`step(res)` として変数選択をさせたりするのに使える。

変数の種類と数の違いによる線型モデルの分類

以下のように整理すると，t 検定，分散分析，回帰分析といった分析法が，すべて一般化線型モデルの枠組みで扱えることがわかる。

分析名	従属変数 (Y)	独立変数 (X)
t 検定（注 1）	量的変数１つ	２値変数１つ
一元配置分散分析	量的変数１つ	カテゴリ変数１つ
多元配置分散分析（注 2）	量的変数１つ	カテゴリ変数複数
（単）回帰分析	量的変数１つ	量的変数１つ
重回帰分析	量的変数１つ	量的変数複数（注 3）
共分散分析	量的変数１つ	（注 4）
ロジスティック回帰分析	２値変数１つ	２値変数，カテゴリ変数，量的変数複数
正準相関分析	量的変数複数	量的変数複数

（注 1） Welch の方法でない場合。

（注 2） 独立変数となるカテゴリ変数（因子）が２つの場合は二元配置分散分析，３つなら三元配置分散分析と呼ばれる。独立変数はカテゴリ変数そのものだけでなく，交互作用項も含めるのが普通である。なお，分散分析をするときには変数ごとに平方和を求めるわけだが，二元配置以上では平方和の求め方が Type I から Type IV まで 4 通りあるので注意が必要である[*4]。

[*4] 分散分析表にでてくる因子の残差平方和の出し方としては，因子が直交していれば（因子間の交互作用がなければ），他の因子を加える順序によらず一定になるので，他の因子を含まない単独のモデルで出した平方和をそのままその因子の平方和とみなしていい（これが逐次平方和と呼ばれる Type I SS）けれど，因子が直交していないときは別の考え方をする必要があって，そこで出てくるのが，Type II とか Type III の平方和である。

Type II は，まずすべての因子の主効果を含むモデルを基準にして，それから１つの因子を取り去ったモデルのモデル平方和と元のモデルのモデル平方和の差を，取り去った因子の寄与とみなして，その因子の偏平方和（Type II SS）とし，次に 2 因子交互作用を含むモデルを基準にして，交互作用を取り去ったモデルのモデル平方和とのモデル平方和の差を交互作用効果の偏平方和とするというもの。

Type III は繰り返し数が不揃いのときにデータ数の少ないセルを他のセルと同等とみなす目的で使うものだが，同等とみなすと逆にバイアスが生じる可能性もあるので，不揃いでも Type II を使うべきという意見もある。Type

（注 3）　カテゴリ変数はダミー変数化せねばならない。

（注 4）　２値変数１つと量的変数１つの場合が多いが，「２値変数またはカテゴリ変数１つまたは複数」と「量的変数１つまたは複数」を両方含めば使える。

こう考えてみると，t 検定は分散分析の特殊な場合ということができるし，分散分析は線型モデルの特殊な場合ということができるし，線型モデルは一般化線型モデルの特殊な場合ということができる。

重回帰分析についての留意点

重回帰分析が独立変数１つの回帰分析よりも優れている点は，複数の独立変数を同時にモデルに投入することにより，従属変数に対する，他の影響を調整した個々の変数の影響をみることができることである。

重回帰分析は，何よりもモデル全体で評価することが大切である。例えば，独立変数が年齢と体重と一日当たりエネルギー摂取量，従属変数が血圧というモデルを立てれば，年齢の偏回帰係数（または偏相関係数または標準化偏回帰係数）は，体重と一日当たりエネルギー摂取量の影響を調整した（取り除いた）後の年齢と血圧の関係を示す値だし，体重の偏回帰係数は年齢と一日当たりエネルギー摂取量の影響を調整した後の体重と血圧の関係を示す値だし，一日当たりエネルギー摂取量の偏回帰係数は，年齢と体重の影響を調整した後の一日当たりエネルギー摂取量と血圧の関係を示す値である。独立変数が年齢と体重で従属変数が血圧である場合の年齢の偏回帰係数は，独立変数に一日当たりエネルギー摂取量も入っている場合の年齢の偏回帰係数とは**異なる**。

モデル全体としてのデータへの当てはまりは，重相関係数の２乗（決定係数）や，AIC で評価する。

あるモデルの中で，各独立変数が他の独立変数の影響を調整した上でも従属変数に有意な影響を与えているかどうかをみるには，独立変数ごとに，偏回帰係数の有意性検定を行う。ある独立変数の偏回帰係数がゼロという帰無仮説を検定するには，その変数と従属変数の間の偏相関係数がゼロという帰無仮説を t 分布を使って検定すればよい。また，１つの重回帰モデルの中で，相対的にどの独立変数が従属変数（の分散）に対して大きな影響を与えているかは，偏相関係数の二乗の大小

IV は SAS には入っているが，あまり使われない。高橋・大橋・芳賀「SAS による実験データの解析」（東大出版会）によると，数量化一類をするときや，乱塊法の場合や， MANOVA の場合や欠損値がある場合は Type II の使用が薦められるとあるので，とりあえず Type I と Type II だけ出せれば充分ではないかと思う。なお，同書の 16 章には，行列言語 IML で Type III を計算する方法が載っている。

R の場合，標準の `anova()` や `aov()` では Type I の平方和が計算されるが（`anova(lm())` が `aov()` と同じ意味），`car` パッケージの `Anova()` では Type II または Type III（後者を出すには type="III" という引数をつける）の平方和が計算できる。ただ，`library(car)` してから `help(Anova)` すると，`Anova()` 関数で計算される Type II は SAS の Type II と同じだが Type III は微妙に違うので注意して使えと書かれている。`car` パッケージの開発者 John Fox の著書 “An R and S-PLUS companion to applied regression.” の p.140 の Type III の説明によると，例えば因子 A の主効果を，因子 B の主効果と因子 A と因子 B の交互作用効果をテストした後でテストしたいような場合に他の効果のすべてを出した後で因子 A によって加えられる分を Type III として計算するとのことである。たしかに SAS の計算アルゴリズムとは違うようである。

結論としては，R で，因子が直交していなくてセルごとの繰り返し数が不揃いの二元配置分散分析をしたいときは，`library(car)` としてから，`Anova(lm(Y~C1*C2))` を使えば Type II の平方和，つまり偏平方和が計算されるので，そうすることをお薦めする。

によって評価するか，または標準化偏回帰係数によって比較することができる。しかし，別の重回帰モデルとの間では，原則として比較不可能である。

多重共線性 (multicolinearity)

一般に，複数の独立変数がある場合の回帰で，独立変数同士に強い相関があると，重回帰の係数推定が不安定になるのでうまくない。ごく単純な例でいえば，従属変数 `Y` に対して独立変数群 `X1` と `X2` が相加的に影響していると考えられる場合，`lm(Y ~ X1+X2)` という重回帰モデルを立てるとしよう。ここで，実は `X1` が `X2` と強い相関をもっているとすると，もし `X1` の標準化偏回帰係数の絶対値が大きければ，`X2` による効果もそちらで説明されてしまうので，`X2` の標準化偏回帰係数の絶対値は小さくなるだろう。まったくの偶然で，その逆のことが起こるかもしれない。従って，係数推定は必然的に不安定になる。この現象は，独立変数群が従属変数に与える線型の効果を共有しているという意味で，多重共線性 (multicolinearity) と呼ばれる。

多重共線性があるかどうかを判定するには，独立変数間の散布図を 1 つずつ描いてみるなど，丁寧な吟味をすることが望ましいが，各々の独立変数を，それ以外の独立変数の従属変数として重回帰分析したときの重相関係数の 2 乗を 1 から引いた値の逆数を VIF（Variance Inflation Factor; 定訳は不明だが，分散増加因子と訳しておく）として，VIF が 10 を超えたら多重共線性を考えねばならないという基準を使う（Armitage et al. 2002）のが簡便である。多重共線性があるときは，拡張期血圧（DBP）と収縮期血圧（SBP）のように本質的に相関するものだったら片方だけを説明変数に使うのが 1 つの対処法である。除かずに調整する方法としては，centring という方法がある。リッジ回帰（R では `MASS` ライブラリの `lm.ridge()`）によっても対処可能である。また，`DAAG` ライブラリ（Maindonald and Braun, 2003）の `vif()` 関数を使えば，自動的に VIF の計算をさせることができる[*5]。

例題 1

第 8 回の課題でも使ったが，`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が使えるようになる。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。
ニューヨーク市のオゾン濃度を，セントラルパークの日照，LaGuardia 空港の平均風速，日最高気温によって説明する重回帰モデルを，このデータに当てはめよ。

重回帰モデルの当てはめと，3 つの独立変数すべてについて Armitage らの方法で VIF の算出を行う R のプログラムは下枠内の通り（なお，`detach()` しない限り `attach()` されたままになり，最初の 2 行を一々入力しなおす必要がなくなる）。`DAAG` ライブラリの使い方は，`res` への付値後に，`require(DAAG)`，`vif(res)` とすれば，3 つの独立変数全ての VIF が得られる。

[*5] 但し Armitage et al. が説明している方法と若干計算方法が異なり，結果も微妙に異なる。

it12-1.R

```
data(airquality)
attach(airquality)
res <- lm(Ozone ~ Solar.R+Wind+Temp)
VIF <- function(X) { 1/(1-summary(X)$r.squared) }
VIF(lm(Solar.R ~ Wind+Temp))
VIF(lm(Wind ~ Solar.R+Temp))
VIF(lm(Temp ~ Solar.R+Wind))
summary(res)
detach(airquality)
```

3 つの独立変数の VIF はすべて 10 より遥かに小さく，多重共線性の問題はないと考えられる。summary(res) の結果は下枠内の通り得られるので，すべての係数が 5%水準でゼロと有意差があり，3 つの独立変数すべてがオゾン濃度に有意に影響しているといえる。また，Adjusted R-squared（自由度調整済み重相関係数の 2 乗）の値から，オゾン濃度のばらつきが，これら 3 つの独立変数のばらつきによって約 60%説明されることがわかる。

```
Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp)

Residuals:
    Min      1Q  Median      3Q     Max
-40.485 -14.219  -3.551  10.097  95.619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Solar.R       0.05982    0.02319   2.580  0.01124 *
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Temp          1.65209    0.25353   6.516 2.42e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.18 on 107 degrees of freedom
Multiple R-Squared: 0.6059,     Adjusted R-squared: 0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

モデルの評価

モデルの当てはめで大事なのは，(1) どのモデルがよりよくデータを説明するのか？　(2) そのモデルはどの程度よくデータを説明しているのか？　を評価することである。以下，簡単にまとめてみる。

線型回帰モデルならば決定係数，すなわち自由度調整済み（重）相関係数の二乗が大きいモデルを採用するというのが 1 つの考え方である。しかし，この基準はかなりナイーブである。一般に，

モデルの採否を決定するための基準としてよく使われるのは，残差分析，尤度比検定，AIC である。

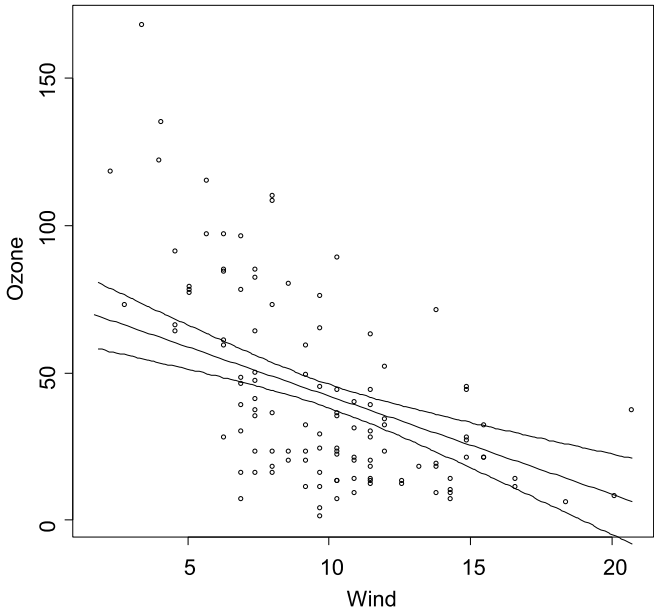
残差分析と信頼区間

残差分析を行うと，モデルがデータから系統的にずれていないかどうかを検討することができる。系統的なズレは，とくにモデルを予測や信頼区間の推定に用いる場合に大きな問題となる（系統的なズレが大きいモデルは，そういう目的には使えない）。回帰モデルの結果を `res` に付値しておけば，例えば，`Wind` の大小と残差の大小の間に関連があるかどうか見るためには，`plot(residuals(res)~res$model$Wind)` とすることで，回帰の結果から残差を取り出してプロットすることができる（横軸としてはすべての独立変数について試してみるべきである）。横軸の大小によらず，縦軸のゼロの近辺の狭い範囲にプロットが集中していれば，残差に一定の傾向がないことになり，系統的なズレはなさそうだと判断できる。なお，横軸の変数を指定せずに，`plot(residuals(res))` したときの横軸は，オブザーベーションの出現順を意味するインデックス値になる。

残差分析の裏返しのようなイメージになるが，信頼区間の推定も有用である。線型モデルであれば，信頼区間の推定には `predict()` 関数を用いることができる。例えば `Wind` のとる範囲に対して 95%信頼区間を得るためには，他の 2 つの変数が平均値で固定されていると仮定して，下枠内のプログラムを用いれば，`Wind` を横軸に，`Ozone` を縦軸にしたデータそのものがプロットされた上で，重回帰モデルによる推定値が実線で，その 95%信頼区間が点線で重ね描きされる。ただし，単回帰分析（独立変数が 1 つだけの回帰分析）の場合ほど意味がクリアではない。

it12-2.R

```
data(airquality)
attach(airquality)
res <- lm(Ozone ~ Solar.R+Wind+Temp)
EW <- seq(min(Wind),max(Wind),len=100)
ES <- rep(mean(Solar.R,na.rm=T),100)
ET <- rep(mean(Temp,na.rm=T),100)
Ozone.EWC <- predict(res,list(Wind=EW,Solar.R=ES,Temp=ET),interval="conf")
plot(Ozone~Wind)
lines(EW,Ozone.EWC[,1],lty=1)
lines(EW,Ozone.EWC[,2],lty=2)
lines(EW,Ozone.EWC[,3],lty=2)
detach(airquality)
```



尤度比検定

次に，モデルの相対的な尤もらしさを考えよう。重回帰分析で独立変数が 3 つの場合とそのうち 1 つを除いた 2 つの場合，あるいは 3 次回帰と 2 次回帰のように，一方が他方を一般化した形になっている場合は，これら 2 つのモデルを比較することができる。

一般に，$f(x,\theta)$ で与えられる確率密度関数からの観測値を $\{x_1, x_2, ..., x_n\}$ とするとき，θ の関数として，$L(\theta) = f(x_1,\theta)f(x_2,\theta)\cdots f(x_n,\theta)$ を考えると，確率密度関数の値が大きいところほど観測されやすいため，$L(\theta)$ の値を最大にするような θ を真の θ の推定値とみなすのが最も尤もらしい。この意味で $L(\theta)$ を尤度関数と呼び，この θ のような推定量のことを最尤推定量と呼ぶ。尤度関数を最大にすることはその対数をとったもの（対数尤度）を最大にすることと同値なので，対数尤度を θ で偏微分した式の値をゼロにするような θ の中から $\ln L(\theta)$ を最大にするものが，最尤推定量となる。例えば，正規分布に従うサンプルデータについて得られる尤度関数を母平均 μ で偏微分したものをゼロとおいた「最尤方程式」を解けば，母平均の最尤推定量が標本平均であることがわかる。詳しくは鈴木 (1995) を参照されたい。

一般に，より一般性の低いモデルをデータに当てはめたときの最大尤度を，より一般的なモデルの最大尤度で割った値の自然対数をとって-2 を掛けた値 λ は，「尤度に差がない」という帰無仮説の下で，自由度 1（比較するモデル間のパラメータ数の差）のカイ二乗分布に従うので，検定ができる。この検定を尤度比検定と呼ぶ。R では，`logLik()` が対数尤度とパラメータ数を計算する関数なので，この関数を使えばよい。

例題 2

例題 1 と同じデータで，独立変数が日照，風速，気温すべてであるモデルと，独立変数が日照と風速だけのモデルを尤度比検定せよ。

下枠内のように入力すれば，尤度比検定した有意確率は 10^{-9} のオーダーなので，有意水準 5%で帰無仮説は棄却される。したがってこの場合は 2 変数よりも 3 変数のモデルを採用すべきである。

it12-3.R

```
data(airquality)
attach(airquality)
res.3 <- lm(Ozone ~ Solar.R+Wind+Temp)
res.2 <- lm(Ozone ~ Solar.R+Wind)
lambda <- -2*(logLik(res.2)-logLik(res.3))
1-pchisq(lambda,1)
detach(airquality)
```

この例題では線型重回帰の関数 `lm()` を扱ったが，この尤度比検定の考え方は，2 つのモデルが包含関係にありさえすれば，一般化線型モデル `glm()` でも非線型モデル `nls()` でも，同じように使える。

AIC: モデルの当てはまりの悪さの指標

さて一方，AIC はパラメータ数と最大尤度からモデルの当てはまりの悪さを表すものとして計算される指標で，数式としては，L を最大尤度，n をパラメータ数として，

$$\mathrm{AIC} = -2\ln L + 2n$$

で表される。AIC が小さなモデルほど当てはまりがいいと考える。

実は，R には，`AIC()` という関数と `extractAIC()` という 2 つの関数がある。前者は “Akaike’s An Information Criterion” となっていて, 後者は “The (generalized) Akaike *A*n *I*nformation *C*riterion for a fitted parametric model” となっている。前者が以前からある汎用関数である。`extractAIC()` は MASS ライブラリに含まれていたのが S4 メソッドとして標準実装されるようになった関数で，変数選択のために `step()` 関数の中から呼び出されるのが主な用途である。

例えば，例題 2 の二つのモデルについて AIC を計算すると，`AIC(res.3)` は，確かに

```
-2*logLik(res.3)+2*attr(logLik(res.3),"df")
```

と同じで 998.7 となり，`extractAIC(res.3)` の結果は 681.7 となる。`res.2` についても同様で，`AIC(res.2)` は 1033.8，`extractAIC(res.2)` は 716.8 を返す。

いずれにせよ独立変数 3 つのモデルの方が AIC が小さく当てはまりがよいとは言えるが，結果が異なるのはおかしいし，困ってしまう。実は，定義通りの AIC を返すのは `AIC()` 関数なのだが，変数選択に使うためならそれと定数の差があってもいいので，計算量が少ない `extractAIC()` 関数が `step()` では使われているということのようである[*6]。

[*6] `http://www.is.titech.ac.jp/~shimo/class/gakubu200409.html`（東工大・下平英寿さんの講義「R による多変量解析入門」の第 8 回「モデル選択」の資料）に，それぞれが使っている式の説明があり，`AIC()` 関数は $-2\ln L + 2\theta$（L は最大尤度，θ はパラメータベクトルの次元）を計算する汎用関数であって，オブザーベーション数 n，パラメータ数 p，標準偏差 σ として，線型重回帰の場合は $n(1+\ln(2\pi\sigma^2))+2(p+1)$ を計算し（正規分布

変数選択

このように，重回帰モデルの独立変数の取捨選択を行うことを変数選択と呼ぶ。非線型モデルの場合は自動的にはできないので，残差分析や尤度比検定や AIC の結果を見ながら手作業でモデリングを進めていくしかないが[*7]，線型重回帰モデルならば，`step(lm(Ozone~Wind+Solar.R+Temp))` のように `step()` 関数を使って，自動的に変数選択を行わせることができる。変数増加法 (`direction="forward"`)，変数減少法 (`direction="backward"`)，変数増減法 (`direction="both"`) などがある。例えば減少法の場合，`direction="backward"`オプションをつけるが，変数選択候補範囲を明示的に与えない場合，`step()` 関数のデフォルトは減少法になっているので，線型重回帰分析の結果を `step()` に渡す場合には，`direction` 指定はしなくても同じ結果になる。

このデータの場合，`ress <- step(lm(Ozone~Solar.R+Wind+Temp))` とすると，3 つすべての変数が残った場合の AIC である 682 が最小であることがわかり，採択されたモデルが `ress` に保存される。ここで表示された AIC は `step()` 関数が，内部的に `extractAIC()` 関数を使って得た値なので，通常の AIC を表示するには，採択されたモデルに対して `AIC(ress)` としなくてはならない。`lm()` で使われたオブザーベーションが 111 しかないので（Ozone と Solar.R に欠損値が多いため），`AIC(ress)-111*(1+log(2*pi))-2` とすると，確かに `681.7127` という結果になり，`step()` 関数の出力に出てくる値と一致することがわかる。まとめると，変数減少法で変数選択をさせ，最終的に採択されたモデルについての情報を表示させるには，下枠内のように入力すればよい。

it12-4.R

```
data(airquality)
attach(airquality)
res <- lm(Ozone~Solar.R+Wind+Temp)
ress <- step(res)
summary(ress)
AIC(ress)
```

重回帰分析では，たくさんの独立変数の候補から比較的少数の独立変数を選択することが良く行われるが，モデル全体で評価するという観点からは，あまり薦められない。数値以外の根拠により投入する変数を決めて，各々の偏回帰係数（または偏相関係数）が有意であるかないかを見る方が筋がよい。十分な理由があれば，有意でない変数も含めた重回帰式を作っても良い。

しかし，数値以外の根拠が薄い場合もあるし，偏回帰係数が有意でない（偏相関係数がゼロであるという帰無仮説が成り立つ確率が 5%より大きい）変数を重回帰モデルに含めることを嫌う立場

を仮定するから），`extractAIC()` 関数は線型重回帰のときだけ使える関数で，$n\ln(\sigma^2)+2p$ を計算する。前者から後者を引けば $n(1+\ln(2\pi))+2$ と，オブザーベーション数は含むけれどもパラメータ数には依存しない定数になるので，変数選択はこちらでやっても問題ないことになる。

[*7] R-help メーリングリストによると，S-plus には `step.glm()` という関数があるらしいが，R では敢えて実装しなかったらしい

もある。従って，数値から最適なモデルを求める必要もありうる。そのためには，独立変数が 1 個の場合，2 個の場合，3 個の場合，……，のそれぞれについてすべての組み合わせの重回帰モデルを試して，最も重相関係数の二乗が大きなモデルを求めて，独立変数が n 個の場合が，$n-1$ 個の場合のすべての変数を含むならば尤度比検定を行って，尤度が有意に大きくならないところまでの $n-1$ 個を独立変数として採用するのが良い。これを総当り法と呼ぶ。M.G. ケンドール著（奥野忠一，大橋靖雄訳）『多変量解析』（培風館，1981）では総当り法が薦められているが，R の `step()` 関数では提供されていない[*8]。

採択されたモデルを使った予測

モデルの当てはめがうまくできれば, 独立変数群の値から従属変数の値を予測することができる。信頼区間の計算で示したように，`predict()` 関数を使えばよい。例えば，風速も日照も気温も観測値の平均値になった日に, オゾン濃度がいくらになるかを, `res<-lm(Ozone~Solar.R+Wind+Temp)` という回帰式から予測するには，上枠内のコードを実行させた後に下枠内を実行する。

it12-5.R

```
predict(res, list(Solar.R=mean(res$model$Solar.R),
 Wind=mean(res$model$Wind), Temp=mean(res$model$Temp))
```

他の観測値がわかっていて，オゾン濃度だけを測れなかった日の値を推定する（補間することになる）のにも，同じ方法が使える。

ただし，第 8 回の資料でも書いたけれども，回帰の外挿には慎重でなければならない。こうして推定された回帰係数を用いて，`Solar.R` と `Temp` がそれぞれこの重回帰分析で使われた値の平均値（なお，重回帰分析で使われた値だけでなく，できるだけ多くの値を使いたい場合なら，Solar.R には欠損値が含まれているので，平均を計算するときに，`mean(Solar.R,na.rm=T)` としなくてはならない）で，`Wind=25` のときのオゾン濃度を点推定すると約`-8.1` となってしまって，やはり採用できない（95%信頼区間はゼロを跨いでいるが）。結局，いくら AIC が小さくなっても，論理的に問題がある線型回帰を適用して予測をしてはいけないということである。

そこで登場するのが非線型回帰である。Wind と Ozone が負の相関関係があるので Wind が大きくなると Ozone がマイナスになるという線型回帰の弱点を避けるために，例えば `Wind` と `Solar.R` の 2 パラメータで，風速が係数が負の指数関数の形でオゾン濃度に影響する非線型関係を仮定したモデルで回帰分析を行うには，下枠内のようにする[*9]。

it12-6.R

```
resmr <- nls(Ozone ~ a*exp(-b*Wind) + c*Solar.R, start=list(a=200,b=0.2,c=1))
summary(resmr)
AIC(resmr)
```

[*8] `http://aoki2.si.gunma-u.ac.jp/R/All_possible_subset_selection.html` に，本学社会情報学部の青木繁伸教授が開発された R コードが公開されている。

[*9] 言うまでもないが，データが使えなくてはいけないので，ここまでの操作に続けて行わねばならない。このコードだけ単独で実行しても無意味である。

AIC は線型の 2 パラメータモデルより小さい（extractAIC() 関数は，非線型モデルには使えない)。独立変数が 1 つだけの場合に比べるとずっと小さい（もっとも，十分に小さいとはいえないので，このデータに含まれていない，他の要因の影響が大きいのであろう)。Temp も入れたモデルと尤度比検定をすると有意ではないので，このモデルが採用できる。そこで続けて下枠内を入力すれば[*10]，

```
SRM <- mean(subset(Solar.R,!is.na(Ozone)&!is.na(Solar.R)&!is.na(Wind)&!is.na(Temp)))
predict(resmr,list(Wind=25,Solar.R=SRM))
```

約 16.4 となるので，風速 25 マイル／時のときのオゾン濃度は，太陽放射が平均的な条件なら，約 16.4 ppb になると予測される。

共分散分析

共分散分析は，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_1 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_2 と相関がある場合に（このとき X_2 を共変量と呼ぶ)，X_2 と Y の回帰直線の傾き (slope) が X_1 の示す 2 群間で差がないときに，X_2 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_1 の 2 群間で差があるかどうかを検定する。

R では，X_1 を示す変数名を C（注：C は factor である必要がある)，X_2 を示す変数名を X とし，Y を示す変数名を Y とすると，summary(lm(Y~C+X)) とすれば，X の影響を調整した上で，C 間で Y の修正平均（調整平均）が等しいという帰無仮説についての検定結果が得られる（C のカテゴリが 1 と 2 である場合，C2 と表示される行の右端に出ているのがその有意確率である)。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。そこで，まず例えば，summary(lm(Y[C==1]~X[C==1]); summary(lm(Y[C==2]~X[C==2]) として 2 つの回帰直線それぞれの適合を確かめ，summary(lm(Y~C+X+C:X))（または summary(lm(Y~C*X))）として傾きが等しいかどうかを確かめなければならない。傾きが有意に違っていることは，C と X の交互作用項が有意に Y に効いていることと同値なので，Coefficients の C2:X と書かれている行の右端を見れば，「傾きに差がない」という帰無仮説の検定の有意確率が得られる。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。

[*10] web 上の it12-6.R には下枠内も含んでいる。

例題 3

第 8 回の例題で使った R の組み込みデータ 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。第 8 回では投与方法の違いを無視して用量と長さの関係を調べたが，用量と長さの関係が投与方法によって異なるかどうかを共分散分析を使って調べよう。

例によってデータを使えるようにしてから，まずグラフを描いてみる。共分散分析をするような場面では，通常，下枠内のように[*11]，群によってマークを変えて散布図を重ね描きし，さらに線種を変えて群ごとの回帰直線を重ね描きするのだが，coplot(len~dose | supp) として横に 2 枚のグラフが並べて描かれるようにすることも可能である。

it12-7.R

```
data(ToothGrowth)
attach(ToothGrowth)
plot(dose,len,pch=as.integer(supp),ylim=c(0,35))
legend(max(dose)-0.5,min(len)+1,levels(supp),pch=c(1,2))
abline(lm1 <- lm(len[supp=='VC']~dose[supp=='VC']))
abline(lm2 <- lm(len[supp=='OJ']~dose[supp=='OJ']),lty=2)
summary(lm1)
summary(lm2)
```

summary(lm1) と summary(lm2) をみると，投与方法別の回帰係数がゼロと有意差があることがわかる。そこで次に，これらの回帰係数間に有意差がないという帰無仮説を検定する。モデルの右辺に独立変数間の交互作用項を含めればいいので，

```
lm3 <- lm(len ~ supp*dose)
summary(lm3)
```

とすると，

[*11] web 上の it12-7.R にはその次の枠内のコードも含む。

```
Call:
lm(formula = len ~ supp * dose)

Residuals:
     Min       1Q   Median       3Q      Max
-8.22643 -2.84625  0.05036  2.28929  7.93857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.550      1.581   7.304 1.09e-09 ***
suppVC        -8.255      2.236  -3.691 0.000507 ***
dose           7.811      1.195   6.534 2.03e-08 ***
suppVC:dose    3.904      1.691   2.309 0.024631 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.083 on 56 degrees of freedom
Multiple R-Squared: 0.7296,     Adjusted R-squared: 0.7151
F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16
```

という出力が得られる。この結果から，suppVC:dose の従属変数 len への効果（交互作用効果）がゼロという帰無仮説の検定の有意確率が 0.024631 となるので，有意水準 5%で帰無仮説は棄却される。従って，この場合は，**投与経路によって投与量と長さの関係の傾きが有意に異なるので，と提示した上で，**先に計算済みの，投与経路別の回帰分析の結果を解釈すればよい（修正平均の差の検定はしても意味がない)。

例題 4

http://phi.med.gunma-u.ac.jp/medstat/sample1.dat は変数名付きのタブ区切りテキスト形式のデータで，5 つの変数が含まれている（PREF が都道府県名，REGION が東日本か西日本か，CAR1990 が 1990 年の 100 世帯当たりの自動車保有台数，TA1989 が 1989 年の人口 10 万人当たり交通事故死者数，DIDP1985 が 1985 年の人口集中地区居住者割合である)。下枠内のコードを実行すると，データフレーム dat に読み込むことができる。

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/sample1.dat")
attach(dat)
```

ここで，東日本と西日本で交通事故死者数に差があるかを検討したいとする。しかし，交通事故死者数は，自動車保有台数と関連がありそうなので，もしそうなら，その影響を調整した上でなお，東日本と西日本で差があるかを検定したい。どうしたらよいか。

まず REGION ごとに自動車保有台数と交通事故死者数の単相関を検討すると，東日本の回帰係数の有意性検定の p-value は 5.72e-05，西日本では 0.00267 と，ともに 5%よりずっと小さいので，有意な関連があるといえる。次に 2 つの回帰直線の傾きに有意差があるかどうかを検定する

と，REGIONWest:CAR1990 の p-value は 0.990 なので，傾きに有意差はないといえる。そこで修正平均の差を検定すると，REGIONWest の係数の検定の有意確率は 0.0319 なので，5%水準では有意である。よって，自動車保有台数を調整しても，東日本と西日本では交通事故死者数に有意差があるといえる。なお，以上を実行するコードは下枠内の通りである（web 上の it12-8.R には読み込み部分も含んでいる)。

it12-8.R

```
plot(CAR1990,TA1989,pch=as.integer(REGION))
legend(max(CAR1990)-10,min(TA1989)+1,levels(REGION),pch=c(1,2))
abline(lm1 <- lm(TA1989[REGION=='East']~CAR1990[REGION=='East']))
abline(lm2 <- lm(TA1989[REGION=='West']~CAR1990[REGION=='West']),lty=2)
summary(lm1)
summary(lm2)
lm3 <- lm(TA1989 ~ REGION*CAR1990)
summary(lm3)
lm4 <- lm(TA1989 ~ REGION+CAR1990)
summary(lm4)
```

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，正規分布に従わないので glm() を使う。

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。

例題として，library(MASS) にある data(birthwt) を使った実行例を示す。

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る。str(birthwt) とすると変数がわかる。

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。

it12-9.R

```
require(MASS)
data(birthwt)
attach(birthwt)
low <- factor(low)
race <- factor(race, labels=c("white","black","other"))
ptd <- factor(ptl>0)
smoke <- (smoke>0)
ht <- (ht>0)
ui <- (ui>0)
ftv <- factor(ftv)
levels(ftv)[-(1:2)] <- "2+"
bw <- data.frame(low,age,lwt,race,smoke,ptd,ht,ui,ftv)
detach(birthwt)
summary(res <- glm(low ~ ., family=binomial, data=bw))
summary(res2 <- step(res))
```

変数選択後の結果をみると，smokeTRUE の係数（対数オッズ比）は 0.866582 で，その SE が 0.404469 である。したがって，最終的なモデルに含まれる他の変数（最終月経時体重，黒人，他の有色人種，非熟練労働経験あり，高血圧既往あり，子宮神経過敏あり）の影響を調整した喫煙の低体重出生への効果（オッズ比とその 95%信頼区間）は，下枠内によって得られる。なお，人種は 3 つのカテゴリがあるので，自動的にダミー変数化されて処理される。

```
exp(0.866582)
exp(0.866582 - qnorm(0.975)*0.404469)
exp(0.866582 + qnorm(0.975)*0.404469)
```

結果はそれぞれ，2.378766，1.076616，5.255847 となるので，喫煙者は非喫煙者に比べて約 2.38 倍（95%信頼区間は [1.08, 5.26]），低体重出生児をもちやすいということを示している（95%信頼

区間の下限が 1 より大きいので，有意水準 5%で有意な影響があったといえる）。

引用文献

- 鈴木義一郎 (1995) 情報量基準による統計解析, 講談社サイエンティフィク.
- Armitage P, Berry G, Matthews JNS (2002) Statistical Methods in Medical Research, 4th ed., Blackwell Publishing.
- Maindonald J, Braun J (2003) Data analysis and graphics using R, Cambridge Univ. Press.

課題

最後の例題のデータ birthwt を使って，児の出生時体重をそれ以外の量的変数によって説明する線型重回帰モデルを作成し，変数選択して結果を解釈せよ。なお，例題実行によって `ftv` をカテゴリ変数として書き換えてしまっているはずなので，`rm(list=ls())` としてメモリ上の変数をすべて消去した後に，`data(birthwt)` から改めて始めること。

結果は学籍番号とともにプリンタ出力して，署名して提出すること。結果の提出をもって出席確認とする。

出席が不安な人のための追加課題

コーヒーの種類｜人	A	B	C	D
グアテマラ	75,80,77	60,62,64	55,57,58	60,61,64
マンデリン	77,75,76	63,62,65	60,57,61	58,59,61
ハワイコナ	78,79,77	61,63,58	60,61,65	62,63,60
インドモンスーン	80,81,76	65,64,62	61,62,60	64,63,61
キューバ	74,73,75	61,58,59	64,58,59	65,62,64

上の表は，4 人の人にコーヒーを飲んでもらってから，単純計算のテストを 3 回ずつ受けてもらったときの得点である（架空のデータである）。このテストの得点について，個人差があるか，コーヒーの種類による影響があるか，人によってコーヒーの種類の得点への影響が異なるか（交互作用効果があるか）を二元配置分散分析せよ。

出席が足りているかどうか不安な人は，この課題への回答を，R のコードを含めて A4 用紙 1 枚以内でまとめ，学籍番号とともにプリンタ出力して署名し，第 13 回の講義時に提出すること。