統計学第13回
「一般化線型モデル入門」

中澤　港
http://phi.ypu.jp/stat.html
<minato@ypu.jp>

一般化線型モデルとは？

- 従属変数群が独立変数群の一次結合と誤差で表されるという形のモデルを線型モデルという（回帰分析はデータへの線型モデルの当てはめである）
- 式で書けば $\boldsymbol{y} = \beta_0 + \beta\boldsymbol{x} + \varepsilon$
- R では glm( ) という関数で実行する。
- glm( ) は量的なデータが正規分布に従う場合だけでなく，二項分布やガンマ分布に従う場合でも，family=binom とか family=gamma と指定することで当てはめが可能（その意味で「一般化」）。変数選択も step 関数で可能。
- 正規分布に従う場合は family=gaussian だが省略可能。出力される統計量は違うが，実は lm( ) と同じ結果。

他の分析をglmの枠組みで捉える

検定手法	従属変数	独立変数
t検定	量的変数1つ	2分変数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	量的変数1つ	カテゴリ変数複数
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	基本は量的変数複数
共分散分析	量的変数1つ	基本は2値変数1つと量的変数1つ
ロジスティック回帰分析	2分変数1つ	2分変数，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

t 検定を線型モデルでやってみる

- テキストには家のタイプの違いで飲料水の硬度の差をみる例を載せたが，ここではもっと簡単な例で練習してみよう。
- 男女5人ずつの身長データがあるとする。

```
sex <- as.factor(c(rep('M',5),rep('F',5)))
height <- c(170,166,182,193,160,155,175,148,166,162)
# グラフを書かせる
stripchart(height~sex)
```

- t 検定には，t.test(height~sex,var.equal=T)
- 得られる結果は summary(lm(height~sex)) と同じ。summary(glm(height~sex)) とした場合は AIC が計算される点が異なるが，得られるモデルそのものは同じ。

重回帰モデルの概念

- 独立変数群が k 個あって X1, X2, ..., Xk とするとき，重回帰モデルは，
  $Y= b00 + b01X1+ b02X2+...+b0kXk+\varepsilon$
  と書ける。
- データは，Y={y1,y2,...,yn} で，
  X1={x11,x12,...,x1n}, ..., Xk={xk1,xk2,...,xkn}
  という形になる。データにあてはめたときの残差平方和（εの二乗和）が最小になるように重回帰モデルの係数（偏回帰係数）b00, b01, ..., b0k を推定するには，最小二乗推定か最尤推定をするのだが，その結果が信頼できるためには，k << n でなくてはならないし，Y, X1, ..., Xk は正規分布に従っている方がよい。

R

重回帰モデルの解釈

- 独立変数群は互いに他の変数の影響を調整する。
- 重相関係数（自由度調整済み）の2乗またはAIC（「赤池の情報量基準」の略で，n log(Q/n)+2(k+1) で得られ [尤度をLとして –2log(L)+2(k+1) と書かれることもある]，モデルの当てはまりの悪さを示す指標）でモデルの当てはまりを評価。
- 各独立変数が従属変数に与える影響は，偏回帰係数（標準化偏回帰係数で相対的な影響の大きさ）と偏相関係数で評価。
- モデル全体として捉えることが重要。
- 変数選択をするなら，ステップワイズ法よりもMAXR法が良いが，ステップワイズ法（lmやglmの結果に対してstep関数を適用）の場合は変数減少法がよいとされる(direction=”backward”)。
- 通常は正規分布を仮定するのでglmでもlmでも同等。glmだとAICが計算されるし，違う分布でも使える。lmなら自由度調整済み重相関係数の2乗（決定係数）が計算されるのが利点。

R

重回帰モデルの適用例

- n が大きくなければ使えないので，R の組み込みデータを使う。data() とすると，組み込みデータの一覧が表示される。data(trees) とすると，31 本のアメリカ桜の倒木の測定値のデータフレーム trees がロードされる
- str(trees) とすると，Girth ，Height ，Volume の3つの変数が含まれていることがわかる。Girth は胸高直径，Height は木の高さ，Volume は体積である。ここで，体積が胸高直径と高さで説明されるモデルを考える。胸高直径と高さにもある程度相関がある（ cor(trees) とするとわかる）のだが，偏相関係数はそれを調整した効果を示す値となる
- **`summary(res <- lm(Volume~Girth+Height, data=trees))`** とすると偏回帰係数，その標準誤差，t 値，有意確率，決定係数（重相関係数の2乗），モデルの F 値と有意確率などが表示される。lm の代わりに glm を使えば AIC も計算される（ただし，AIC(res) とすれば lm でも AIC は出せる）。切片がゼロのモデルを当てはめるには，モデルの右辺に 0+ という項を入れておく。
- 次に，**`sdd <- c(0,sd(trees$Girth),sd(trees$Height))`** として各独立変数の不偏標準偏差ベクトルを作り（ 0 は切片用），
  `stb <- coef(res)*sdd/sd(trees$Volume)` とすれば，stb に標準化偏回帰係数のベクトルが得られる。重回帰分析の結果としては，通常，上記のような値を表の形で示すことになっている。

共分散分析のモデル

- 典型的には，従属変数を2分変数と共変量とその交互作用で説明するモデル，即ち X1 を2分変数，X2 を共変量とし，$Y = b0+b1X1+b2X2+b12X1X2+\varepsilon$
- 考え方としては，2分変数によって示される2群間で従属変数の平均値に差があるかどうかを見たい場合に，従属変数の値が共変量によっても影響を受けているなら，その影響を調整しなくてはならないということ
- 共変量と従属変数の回帰直線の傾きが2分変数によって示される2群間で異なるかどうかということと，共変量の影響を調整した従属変数の平均値（調整平均とか修正平均という）が2群間で異なるかどうかを見る。テキストでは glm( ) を使って分析しているが，従属変数が正規分布に従うなら lm( ) でいい
- 図は散布図に2本の回帰直線を重ね書きするのが普通

共分散分析の例

- テキスト p.148-149 の例題のデータを表計算ソフトで入力してからタブ区切りテキスト形式で保存するか，インターネットに繋がった環境なら，

```
x <- read.delim("http://phi.ypu.jp/statlib/l14-1.dat")
```

- 知りたいのは，東日本と西日本で人口集中地区居住割合 (DIDP1985) と 100 世帯当たり乗用車台数 (CAR1990) の関係が異なるか？　関係に差がなければ，DIDP1985 の影響を調整しても CAR1990 に東西で差があるか？
- REGION 別に summary(lm(CAR1990~DIDP1985,x)) を見ると，どちらも回帰係数はゼロとはいえないので，次は summary(lm(CAR1990~factor(REGION)*DIDP1985,x)) によって交互作用項 factor(REGION)2:DIDP1985 の係数がゼロであるという帰無仮説が棄却できないことから傾きに差はないとみて，summary(lm(CAR1990~factor(REGION)+DIDP1985,x)) で factor(REGION)2 の有意確率をみて修正平均の差を検討する。

ロジスティック回帰分析のモデル

- 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分変数であり，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）
- 疾病の有病割合を P とすると，
  log(P/(1-P))=b0+b1X1+...bkXk と定式化できる。X1 が要因の有無を示す2値変数で，X2,...Xk が交絡であるとき，
  X1=0 の場合を X1=1 の場合から引けば，
  b1=log(P1/(1-P1))-log(P0/(1-P0))
  =log(P1*(1-P0)/(P0*(1-P1)))
  となるので，b1 が他の変数の影響を調整したオッズ比の対数になり，オッズ比の 95% 信頼区間が
  exp(b1±1.96*SE(b1)) として得られることがわかる

R

ロジスティック回帰分析の例

- library(MASS) にある data(birthwt) を使ってみる。
  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，　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：児の出生児体重 (kg)

- attach(birthwt) してから

```
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+"
#  必要な変数だけのデータフレーム bwt を定義する
bwt<-data.frame(low,age,lwt,race,smoke,ptd,ht,ui,ftv)
detach(birthwt)
summary(res<-glm(low ~ ., family=binomial, data=bwt))
#  変数選択するなら
summary(res2<-step(res))
```

- 変数選択した場合の結果から，smokeTRUE の係数は 0.866582 で，その SE が 0.404469 なので，他の変数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，exp(0.866582)，exp(0.866582 - 1.96*0.404469)，exp(0.866582 + 1.96*0.404469)，つまり 2.378766，1.076616，5.255847（通常は 2.38 [1.08,5.26] のように表記する）である。喫煙者は非喫煙者に比べて 2.38 倍，低体重出生児をもちやすい傾向にあるといえる。

一般化線型混合モデル

- これは，一般化線型モデルよりも，さらに一般的なモデルである。なぜかというと，個体ごとの経時的変化に代表される，ランダムなばらつきとしての個体差をモデルに取り込めるから（逆にいえば，個体差や部分集団による差の影響を何らかの分布をもった定数項に吸収させることによって除去できる）
- SAS では PROC MIXED で実行できる。
  R では nlme ライブラリの lme( ) 関数を使うか（但し正規分布を仮定できる場合），MASS ライブラリの glmmPQL( ) 関数を使うか，glmmML ライブラリの glmmML( ) 関数を使えば可能（難しいので説明は省略）