前回のQ＆A

Q1) 図を黒板に書くとき，あとから見て何の図か分かるように，図の名前を黒板に書いてください。重要なところは何かわかる印（☆など）をして欲しい。

A1) わかりました。できるだけそうするよう努力します。

Q2) 試験はプリントなど持ち込み可で電卓がいるようになっていますけど，確かずっと前に電卓を持っている人が少ないから電卓はいらないって言いませんでしたか？　ないなら買わないといけないですか？　プリント持ち込み可であるけど，どのようなテスト勉強をすればいいのですか？

A2) 電卓は持ち込み可ですが，なくても構いません。計算は筆算で簡単にできる程度のものしか出題しません。ただ，極端なことをいえば，コンピュータ持ち込み可でもいいくらいだと思っています。なぜなら，統計学の講義の目的は，データを前にして自力で正しい統計解析ができたり，本や論文で使われている統計解析を見てその意味を正当に判断できるようになることだからです。他人の答えをみたり他人に尋ねるのは，自力でないのでダメですが，本やプリントを参照して正しい判断ができれば，統計学の考え方は十分に身についていることになると思います。テストのための勉強としては，配布資料を最初から読み直してみてください。この話題はプリントのあの辺にあったな，という勘が働くようになれば，大抵の問題には答えられるでしょう。

* なお，来週で最終回なので，復習をしていて疑問に感じたことがあれば，A４サイズの紙１枚以内に書いて講義の前に提出するか，あるいは minato@ypu.jp 宛てに，来週の講義開始の 30 分前までにメールで尋ねてください。来週の予定の講義終了後に，できるだけそれらの質問に答えるつもりです。

統計学第 13 回　一般化線型モデル入門

（１）一般化線型モデルとは？

- 第 11 回に説明したように，普通の量的変数の間の線型回帰を一般化すれば，t 検定，分散分析，共分散分析，回帰分析，重回帰分析，ロジスティック回帰分析，正準相関分析など多くの分析方法を共通の数学モデルで扱うことができる。このモデルは一般化線型モデル（または一般線型モデル）と呼ばれる。英語では Generalized Linear Model または General Linear Model といい，SAS でのプロシージャ名も `PROC GLM` だし，R での関数名も `glm` である[1] 。
- 一般化線型モデルは，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，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）。係数は最小二乗法または最尤法で数値的に求める。以下，先にあげたいくつかの分析が，どのように一般化線型モデルを特殊化したものなのかを説明し，その中で重回帰分析と共分散分析について若干の補足説明を加える。

（２）変数の種類と数の違いによる分類

- 以下のように整理すると，これら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	従属変数の種類と数 (Y)	独立変数の種類と数 (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	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	量的変数複数（注２）
共分散分析	量的変数１つ	（注３）
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）カテゴリ変数はダミー変数化
（注３）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。

- 例えば，建物の型の変数 (BD) を集合住宅１，一戸建て２とした場合の，東京のとある大学の学生実習で測定した水道水質の総硬度 (HARD) の平均値に，建物の型によって差があるかどうかを検定したいとする。

[1] 線型は linear の訳で，一次結合という意味なのだが，漢字としては線形と書かれることもある。厳密な区分はないように思われるが，GLM の場合は「型」の字を使う方が普通のようである。

- 等分散性を仮定すれば，R では，
  `BD _ c(1,1,1,1,1,1,2,2,1,1,2,1,1,2,1,1,2,2,1,1,1,1,1,2,1,1,2,1,1,2,1,2,1,1);`
  `HARD _ c(88.280, 103.500, 119.600, 96.210, 109.340, 100.500, 81.390, 75.715,`
  `112.880, 101.150, 84.400, 102.900, 65.000, 97.445, 101.850, 79.100, 103.620,`
  `69.270, 97.090, 101.150, 89.820, 108.560, 98.810, 103.620, 85.940, 89.230,`
  `69.300, 101.150, 101.150, 73.070, 62.695, 148.590, 93.080, 103.500);`
  `t.test(HARD ~ BD, var.equal=T)` とすることによって，以下の結果が得られる。
  Two Sample t-test
  data: HARD by BD
  t = 0.8843, df = 32, p-value = 0.3831
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
  -7.444719 18.867802
  sample estimates:
  mean in group 1 mean in group 2
  96.35354 90.64200
- 一般化線型モデルを使って，建物の型を独立変数として総硬度を従属変数としたモデルの当てはめをしてみると，R では，`summary(glm(HARD ~ BD))` によって，
  Call:
  glm(formula = HARD ~ BD)
  Deviance Residuals:
  Min 1Q Median 3Q Max
  -33.659 -8.957 3.301 7.061 57.948
  Coefficients:
  Estimate Std. Error t value Pr(>|t|)
  (Intercept) 102.065 8.861 11.518 6.41e-13 ***
  BD -5.712 6.459 -0.884 0.383
  —
  Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  (Dispersion parameter for gaussian family taken to be 294.4717)
  Null deviance: 9653.4 on 33 degrees of freedom
  Residual deviance: 9423.1 on 32 degrees of freedom
  AIC: 293.72
  Number of Fisher Scoring iterations: 2 が得られる。Coefficients:の BD のところを見ると，t value が -0.884，その有意確率が 0.383 となっていて，t 検定の結果と一致している（t 値の符号が違うが，t 分布は左右対称なので両側検定では符号が違っても同じ意味）ことがわかる。
- この場合は，当然のことながら，普通の線型モデルでも同じ結果が得られるし，分散分析でも同じ結果となる。つまり，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（当然だが）ということができる。

（3）重回帰分析

- 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。
- モデル全体で評価することが大切。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。
- 偏回帰係数の有意性検定は，偏相関係数がゼロである確率を t 検定によって求める。1 つの重回帰式の中で，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，原則としては，別の重回帰モデルとの間では比較不可能である。

- たくさんの独立変数の候補からステップワイズ法によって比較的少数の独立変数を選択することが良く行われる。しかし，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。SAS では PROC REG の MAXR というオプションで可能である。

（４）共分散分析

- 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群間で差があるかどうかを検定した上で，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。
- 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，R では `summary(glm(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（`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，本当は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。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。

（※）補足：一般化線型混合モデル

- 複数の対象についての経時的観察データが複数あるときに，個体間の経時的な変化のパタンの違いをモデルに取り込むことによって一般化線型モデルをさらに一般化したのが一般線型混合モデル (General Linear Mixed Model) である。R では，nlme というライブラリが提供されている。８歳から 14 歳まで２年おきに歯列矯正の指標として，頭蓋の X 線写真により下垂体から翼上顎裂までの距離 (mm) を，男児 16 人，女児 11 人について測定したデータ（Orthodont という組み込みデータ）による実行例は，`library(nlme); example(lme)` とすれば見ることができる。年齢によるモデル，性と年齢と個体差によるモデルについて出力される。