前回のQ＆A

Q1) 共分散分析もう１度お願いします。
A1) わかりました。
Q2) どのような形式でテストを行いますか？(穴埋めか文章で答えるか計算とか)
A2) そのすべてです。

統計学第 14 回　高度な解析法についての概説

（１）共分散分析

- glm の枠組みで説明すると，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。
- 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(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，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(glm(Y ~ C+X+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,「傾きが等しい」を帰無仮説とした場合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。
- なお，glm としてでなく計算するための数式も書いておく。今，`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，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。切片が等しいという帰無仮説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) わかりにくいと思うので，例題で説明する。下の表は，都道府県別のデータで，1990 年の 100 世帯あたり乗用車台数（CAR1990) と，1989 年の人口 10 万人当たり交通事故死者数 (TA1989) と，1985

年の国勢調査による人口集中地区[1] 居住割合 (DIDP1985) である。REGION の 1 は東日本，2 は西日本を意味する。

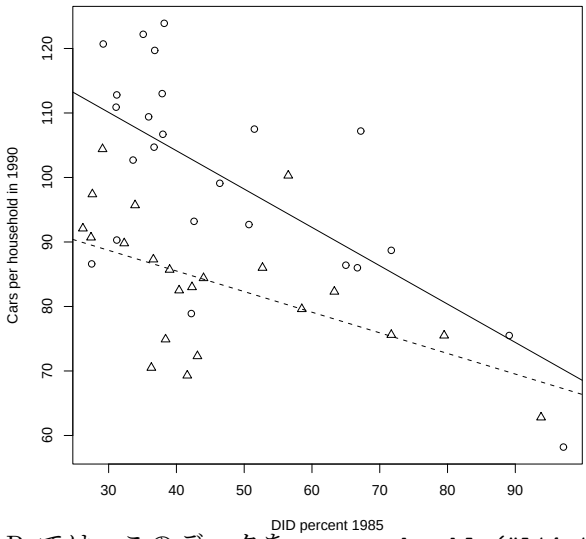
PREF	REGION	CAR1990	TA1989	DIDP1985
Hokkaido	1	86	11.6	66.7
Aomori	1	78.9	9.5	42.2
Iwate	1	86.6	9.7	27.5
Miyagi	1	92.7	7.9	50.7
Akita	1	90.3	8.1	31.2
Yamagata	1	104.7	7.1	36.7
Fukushima	1	102.7	12.1	33.6
Ibaraki	1	120.7	16.4	29.2
Tochigi	1	122.2	16.5	35.1
Gunma	1	123.9	11.5	38.2
Saitama	1	88.7	7.3	71.7
Chiba	1	86.4	8.8	65
Tokyo	1	58.2	4.1	97.1
Kanagawa	1	75.5	7.2	89.1
Niigata	1	93.2	11.1	42.6
Toyama	1	113	11.1	37.9
Ishikawa	1	99.1	9.5	46.4
Fukui	1	109.4	14.7	35.9
Yamanashi	1	112.8	13.8	31.2
Nagano	1	110.9	9.6	31.1
Gifu	1	119.7	12	36.8
Shizuoka	1	107.5	10.5	51.5
Aichi	1	107.2	8.2	67.2
Mie	1	106.7	13.7	38
Shiga	2	104.4	14.5	29.1
Kyoto	2	75.5	8.9	79.5
Osaka	2	62.8	5.9	93.8
Hyogo	2	75.6	8.9	71.7
Nara	2	86	9.3	52.7
Wakayama	2	83	11.6	42.3
Tottori	2	92.1	11.8	26.2
Shimane	2	86.9	9.9	23.4
Okayama	2	95.7	11.3	33.9
Hiroshima	2	79.6	9.7	58.5
Yamaguchi	2	84.4	11.5	44
Tokushima	2	90.7	10.9	27.4
Kagawa	2	89.8	14.3	32.3
Ehime	2	72.3	10.9	43.1
Kochi	2	74.9	11.3	38.4
Fukuoka	2	82.3	8	63.3
Saga	2	97.4	12.8	27.6
Nagasaki	2	69.3	5.9	41.6
Kumamoto	2	87.3	8.5	36.6
Oita	2	82.5	8.7	40.4
Miyazaki	2	85.7	7.4	39
Kagoshima	2	70.5	7.3	36.3
Okinawa	2	100.3	7.6	56.5

- 人口集中地区人口割合が高い都道府県ほど人がまとまって住んでいるわけだから，先験的に，そういう都道府県ほどマイカー保有率は低くて済みそうだと思われる。そこで，この 2 者間に関連があるかどうかを調べてみる。ただし，公共交通機関の整備の割合や，自動車産業の発達の度合い，ディーラーの営業活動の熱心さ，平均世帯規模，郊外型大型店舗の展開の度合い，道路政策，等々，この両者の関係に影響しそうな要因は多々ある。経験からすると，群馬県や長野県は，山口県や熊本県

[1] 1 km^2 当たりの人口密度が 4,000 人以上の集合地区で，かつ合計人口が 5,000 人以上の地区。

に比べて，車が多いような気がしたので，ここで仮に，東日本は，西日本よりも，車をもつ傾向が高いのではないかという仮説を検討してみることにする。

- 東日本を○で，西日本を△でプロットし，東日本の回帰直線を実線，西日本の回帰直線を点線で追加すると，下図のようになる。



- R では，このデータを x _ read.table("l14-1.dat"); attach(x); として読み込み，まず
summary(lm(CAR1990[REGION==1] ~ DIDP1985[REGION==1])),
summary(lm(CAR1990[REGION==2] ~ DIDP1985[REGION==2]))
とすれば，これらの回帰式が，ともに有意にデータに適合していることがわかる。次に，
summary(glm(CAR1990 ~ as.factor(REGION)+DIDP1985+as.factor(REGION)*DIDP1985))
とすれば交互作用項の係数のが有意であるかどうかをみることができて，この結果では確率が 0.118 なので傾きには差がないとわかる。最後に
summary(glm(CAR1990 ~ as.factor(REGION)+DIDP1985)) として as.factor(REGION)2 の有意確率をみると 0.05 より遥かに小さいので，修正平均にも差があるとわかる。つまり，東日本と西日本では，人口集中地区への居住割合の影響を調整しても，世帯当たりの自動車保有台数には有意に差があるといえる。

（2）主成分分析

- n 個体のサンプルがあって，それぞれについて，p 個の変数 $x_1, x_2, ..., x_p$ の観測値が得られているとする。一般に，p 個の変数の情報を全部一度に考えて n 個体の情報を把握することは難しい。そこで考えられるのが，p 個の変数を，もっと少ない数の，互いに独立な主成分 (principal component) で表せないかということである。
- いま，主成分 $\xi_1, \xi_2, ..., \xi_p$ を考え，これらを x の一次関数で表すことにする。つまり，

$$\xi_i = \sum_{j=1}^{p} l_{ij} x_j$$

として，p^2 個の適当な係数 l_{ij} を見つけることを考える。各 x_j をそれぞれの平均からの偏差として測れば，どの x_j も n 個体についての和はゼロになり，従って ξ_i の和もゼロになる。ここで p 個の ξ は互いに無相関であるとする。すなわち

$$E(\xi_i \xi_j) = E(\{\sum_{k=1}^{p} l_{ik} x_k \sum_{m=1}^{p} l_{jm} x_m\} = 0, i \neq j$$

とする。これだけではまだ $p(p+1)/2$ 個の自由度が残っているので，この変換を直交変換であると条件付ける，すなわち

$$\sum_{k=1}^{p} l_{ik} l_{jk} = 0(i \neq j), = 1(i = j)$$

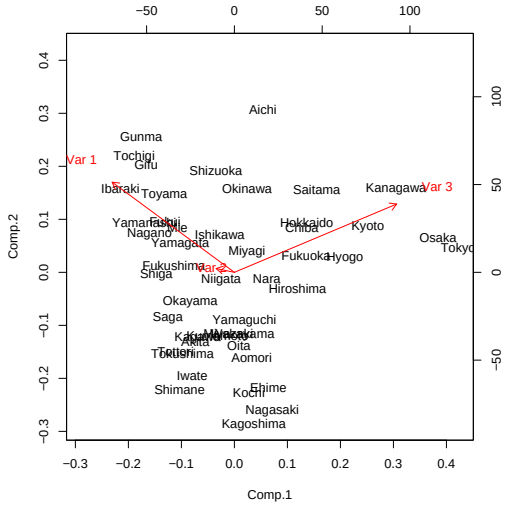
とすれば，符号の付け替えの自由度を加味しても有限組の解が得られることになる（数学的な解は行列の固有値と固有ベクトルを求めることによって得られるが，普通はコンピュータソフトにやら

せるので説明は省略する)。より詳しくは，M.G. ケンドール著（大橋靖雄・奥野忠一訳)「多変量解析」(培風館）を読まれることをお薦めする。

- この新しい変数 ξ は主成分と呼ばれる。もとの変数 x が正規分布に従うなら，ξ は互いに無相関かつ互いに独立である。行列の固有値の大きさの順に $\xi_1, \xi_2, ..., \xi_p$ と番号をつけると，これらは順に第 1 主成分，第 2 主成分，...，第 p 主成分と呼ばれる。第 1 主成分は，あらゆる一次関数の中で可能な最大の分散をもつ。第 2 主成分は第 1 主成分と無相関な一次関数の中で可能な最大の分散をもつ。このようにして主成分を決めると，それぞれの固有値の，固有値の和に対する割合を使って，それぞれの主成分が全変動の何パーセントを説明するかを表すことができる。それを主成分の寄与率と呼ぶ。普通は，たくさんの変数から少数（例えば 2 つとか 3 つ）の主成分だけを使って全変動の 80%が説明できる，のように使う。
- 本当はこんなに少数のデータに使うような分析法ではないのだが，上の例題で使ったデータについて，R を使って主成分分析をしてみる。まず，`library(mva)` として多変量解析ライブラリを呼び出しておく必要がある。ついで，`mat _ matrix(c(CAR1990,TA1989,DIDP1985),nrow=47)` として `res _ princomp(mat); summary(res)` とすれば，下表が得られる。

	Comp.1	Comp.2	Comp.3
Standard deviation	21.1842224	11.7510982	1.897637799
Proportion of Variance	0.7600359	0.2338654	0.006098678
Cumulative Proportion	0.7600359	0.9939013	1.000000000

- この結果から，第 1 主成分の寄与率が 76%，第 2 主成分までの累積寄与率が 99%で，取り上げた 3 つの変数のばらつきは，ほぼ完全に 2 つの直交する主成分に分解できることがわかった。そこで，各都道府県の第 1 主成分（得点）と第 2 主成分（得点）を図示するには，`biplot(res,xlabs=PREF)` とすれば，下図が得られる。

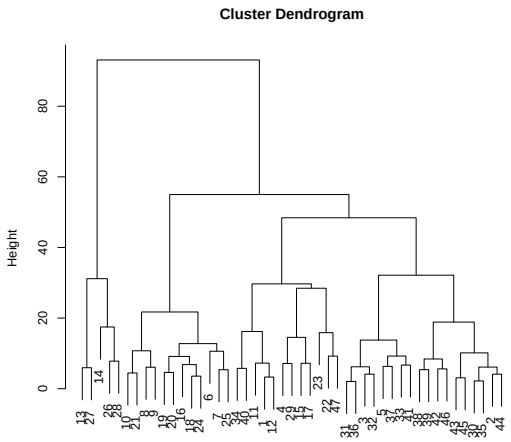


- 思想は逆だが，数学的には因子分析は，主成分分析に良く似ている。つまり，p 個の観測された変数 x があるときに，個々の x が m 個（$m < p$）の潜在因子の線型結合と誤差によって表されると考える。たくさんの変数を，別の少数の変数の線型結合によって表すことによって情報を集約する方法論である。R では `factanal` という最尤法で因子分析を行う関数があるが，3 つの変数を 2 つの因子で説明することはできず，1 つの因子しか想定できない（上の例題のデータで `factanal(mat,2)` とするとエラーが出る。)。`factanal(mat,1)` とすると，第 1 因子の因子負荷量は 1.764 であり，寄与率は 0.588 である。このことは，取り上げた 3 つの変数は，共通の潜在因子によって約 59%説明されるということを意味する。少数の主成分また因子による累積寄与率を最大にするために varimax 回転や promax 回転を行うことがあるが，R ではこれらの関数も用意されている。
- 因子分析は，観測された変数 (observed variables; 観測変数) 間の関係が，実は測定不可能な構成概念 (construct)，即ち因子 (factor) との関係によって説明されると捉えるモデルであるということもできる。しかし因子分析には，観測変数間の関係は，因子との関係においてしか説明できないし，因子間の因果関係を論じることができないし，仮説検証ができないという欠点がある。そこで，測定不可能な因子間の関係もあるだろうけれど，すべてをそれで説明しようとするのではなくて，観測変数間の直接的な関係をまず考えて，それで説明しきれない部分を測定不可能な潜在変数 (latent variables) と変数間の因果関係を不完全にする偶然変動としての誤差変数 (error variables) によって

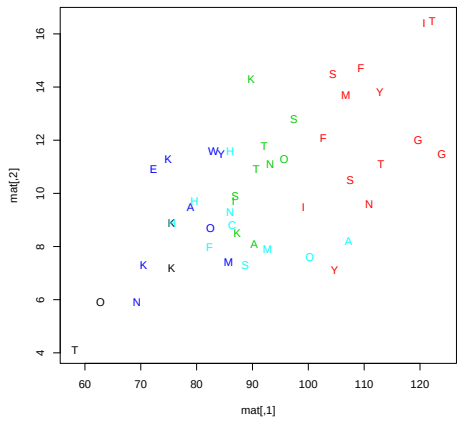
補う，というアプローチが考えられる。これが共分散構造分析 (covariance structure analysis)[2] である。その前段階として，因子分析に仮説検証機能を追加した確認的因子分析がある。共分散構造分析は，潜在変数間の関係を表す構造方程式モデルと，観測変数間の関係を表す測定方程式モデルを，誤差変数を入れて結合したものであるということができる。統計パッケージでは，SAS では PROC CALIS で実行するし，SPSS では AMOS という追加パッケージを使う。R でも sem というライブラリで実行できる。

（3）クラスター分析

- 変数間のでなく，データ間の関係を表したいときに使うのがクラスター分析である。クラスター分析には，距離行列に基づいて個体を結合しながらクラスターを積み上げていく（出力は樹状図またはネットワーク図になる）階層的手法と，予めいくつくらいの塊（クラスター）に分かれるかを決めて，データを適当に振り分ける非階層的手法がある。
- 距離行列の計算法にも多々あり，結合法にも多々ある。いくつかの方法でやってみて，樹状図に差がなければ，そのクラスター分析の結果は安定していて，信頼できるといえる。樹状図が大きく変わるようなら信頼できない。解釈としては，変数が足りないために，個体間の関係が十分にわからないと考える。例としては，R で，先ほどのデータを読んで mva ライブラリを呼び出した後で，`mat _ matrix(c(CAR1990,TA1989,DIDP1985),nrow=47); dis _ dist(mat,method="euclidean"); clus _ hclust(dis); plot(clus,xlab=PREF)` とすれば，樹状図 (dendrogram とか tree とかいう）が下図のように描ける。R ではデフォルトの距離の計算法はユークリッド距離（要するに差の二乗和），クラスター結合法は，完全連結法 (complete linkage) である。クラスター分析の結果は見やすいが，解釈には主観が入りがちである（ちなみに下図で山口は 35 番。30 番の和歌山と最も近いようである）。



- 非階層的手法の k-means 法の R での実行例も示しておく。 5 つのクラスタを仮定すると，`clus5 _ kmeans(mat,5); plot(mat,col=clus5$cluster)` によって下図が得られる。



[2] 共分散構造解析と訳すこともある。