

「データ解析」（下平英寿）　講義資料１

イントロダクション

- 講義のあらまし
- R の簡単な説明
- 簡単なデータ解析の例
- 卒論の紹介
- 講義で例題に用いるデータセットの説明
- 課題

1　講義のあらまし

1.1　「データ解析」の講義について

- Rを用いた多変量解析入門
- 目標
  1. Rを利用して実践的なデータ解析ができるようになること（高度なデータ解析手法）
     ⇒ ライブラリに含まれる関数を呼び出してデータ解析を実行する
  2. 背後にある数学，統計学，アルゴリズムを理解すること（基本的なデータ解析手法）
     ⇒ 自分自身で関数を記述し，それを用いてデータ解析を行う
- R プログラミング技法に関しては,「習うより慣れろ」とする．文法などの講義は行わず,例題を通して実践的に解説する． 文法は C や Java と似ているのですぐに分かるだろう．それよりも，膨大なライブラリ群を把握することが困難かもしれない．
- ホームページ：`http://www.is.titech.ac.jp/~shimo/class/` から本年度の「データ解析」のリンクがある．講義資料の PDF は毎回ホームページに掲載されるので，各自必要に応じて参照・ダウンロードすること．とりあえず昨年度の PDF があり，それを適宜修正，変更していく．
- 成績評価：主にレポート提出でおこなう．期末テストは行わない．
- 出席：成績に反映される．
- レポート課題について
  1. 端末室（もしくは自宅のパソコンなど）で R を用いて課題に取り組みレポートを作成する．
  2. 比較的容易な課題を多く出す．(＊のついた課題は若干難しいもの．)

3. 質問メール受付は shimo-data@is.titech.ac.jp まで遠慮なくメールしてください. 基本的にティーチングアシスタントの大学院生が答えます.

- レポート提出について

1. メールで PDF ファイルを提出. shimo-data@is.titech.ac.jp へ送る. 課題毎にひとつの PDF ファイルとし, プログラム等のテキストもその PDF に含めること.
2. 学籍番号と課題名をファイル名に含める. たとえば学籍番号 0312345, 課題 1 ならば, 0312345-1.pdf とする. もし課題１の小問題を複数のファイルで提出する場合などは, 0312345-1a.pdf, 0312345-1b.pdf のように, ファイル名を変える.
3. レポートの表紙には, 科目名（データ解析）, 氏名, 学籍番号, 課題番号を明記してください.
4. 提出締め切りは, 課題が出されてから原則として１ヶ月以内. 最終レポート締め切りは 8 月の予定です. 詳細はウェブに載せます.
5. やむを得ず紙で提出する場合は, 印刷したものを西８ W の３階にある「下平」のレポートボックスへ提出するか, 直接下平へ手渡す.

1.2 講義予定

確率変数, ロバスト統計, 回帰分析, モデル選択, ブートストラップ法, 主成分分析, 判別分析　など, 多変量解析の入門的内容.

1.3 R を用いた統計解析：参考文献

- R の前身である S 言語を用いたデータ解析入門の教科書：渋谷政昭＋柴田里程　（１９９２）　「S によるデータ解析」　共立出版.
- R（もしくは S) を用いた統計解析の定番教科書：W. N. Venables, Brian D. Ripley 著, Modern Applied Statistics with S 第４版, Springer-Verlag, ２０００年.（ただし統計に関するある程度の知識を仮定しているという意味では少し高度な内容. この本の日本語訳本もあるので, そちらを参照しても良いだろう.）
- R を用いた統計解析の入門書：間瀬茂 他 著,「工学のためのデータサイエンス入門―フリーな統計環境 R を用いたデータ解析」, 数理工学社, ２００４年.（当専攻の間瀬茂先生らによる著書. とても分かりやすいという評判が高い.）
- R の活用事例集：岡田昌史 (編) 「The R Book―データ解析環境 R の活用事例集」 九天社, ２００４年.（R がどのように使われるかを眺めるのに参考になるようです.）

1.4 多変量解析：参考文献

- とくに回帰分析に関しては：佐和隆光　（１９７９）　「回帰分析」　朝倉書店.
  は理論的にも十分詳しく書かれている.

- 多変量解析の教科書は非常に多く出版されているが，例えば：柳井晴夫，高根芳雄　（１９８５）　「新版　多変量解析法」　朝倉書店.

1.5　R の参考情報

- 本家サイト： The R Project for Statistical Computing (`http://cran.r-project.org`)
- 日本語による Wiki 情報サイト：RjpWiki (`http://www.okada.jp.org/RWiki/`) R の情報交換の場になっているようです.

- R の日本語マニュアル：当専攻の間瀬茂先生のページにマニュアルの日本語訳があります (`http://www.is.titech.ac.jp/~mase/R.html`)．PDF 版は東京学芸大学の森厚さんページ (`http://buran.u-gakugei.ac.jp/~mori/LEARN/R/`)．特に，R の「公式マニュアル日本語版 Introduction to R ver.1.7.0」を入手して，ざっと目を通すことをお勧めします．Appendix A の「入門セッション」をとりあえず実行してみるのも良いでしょう.

2　R の簡単な説明

2.1　R とは？

- データ操作，統計計算，グラフィックスのための統合ソフトウエア環境.
- 行列操作に優れている，データ解析の一貫したツール群，簡単で効率的なプログラム言語.
- R はフリーソフトでソースも公開 (`http://cran.r-project.org`).

- Rの開発は１９９０年代後半からネット上で行われている．安定して広く使われるようになったのは２０００年ころから．
- Rの前身であるはSはC言語やUNIXと同じAT&T(現Lucent Technologies) のベル研究所で１９８４年ころ開発 (ちなみにC言語およびUNIXの開発は１９７１年ころ)．
- 現在では，膨大なライブラリがユーザによって開発されている．
- そのほかの統計関連ソフトウエア： SAS, SPSS, Mathematica

2.2 Rの利用

起動 OSのコマンドラインから `% R [return]`

終了 Rのコマンドラインから `> q() [return]` のあとに
`Save workspace image? [y/n/c]:` に対してyと打つ．これで作業ディレクトリに`.RData`というファイルが自動的に作られて定義したオブジェクトが保存される．次回Rを起動したときに自動的に読み込まれる．以降，Rのコマンドラインからの入力を>によって示す．

代入 `> a <- 1:10` は (1,2,...,10) というベクトルを`a`に代入．

計算 `> a^2` は`a`の要素を２乗して結果を表示

```
 [1]   1   4   9  16  25  36  49  64  81 100
```

グラフ `> plot(a,a^2)` は`a`と`a^2`の２次元プロット（散布図と言う）．

関数定義 `> foo <- function(x) sum(x^2)` は要素の２乗和を求める関数を定義し`foo`に代入．呼び出しは`foo(a)`とすれば，`[1] 385`と結果が表示される．

繰り返し `for(i in 1:10) {...}` は`i`を1,...,10まで変化させて括弧内を実行．

```
> x <- rep(0,10); for(i in 1:10) x[i] <- i^2
> x
 [1]   1   4   9  16  25  36  49  64  81 100
```

ヘルプ `> help(for)` は`for`文（と関連する制御構造）についての解説．`> help(":")` は:オペレータの解説．

ライブラリ `> library()` はシステムにインストールされているライブラリパッケージの一覧表示．`> library(MASS)` はMASSライブラリをロード．

デモ `> demo()` でデモの一覧．たとえば`> demo(graphics)`や`> demo(image)`等で`[return]`を押していけばグラフのデモが見れる．

emacsユーザ はESSというemacsパッケージを利用すると便利．(M-x R でRを起動する．)

講義で用いるデータファイル等 は講義ホームページ
`http://www.is.titech.ac.jp/~shimo/class/`におきます．

3　２変量の関係を調べる

3.1　データ

dat0001: 日本の 47 都道府県について，２変量（「学歴」と「出生率」）の値．サイズ 47×2 の実数行列．後述する X2000 データセットの一部であり,「コード」は X2000 における変数のコード．変量 (variate) と変数 (variable) をほぼ同じ意味に用いている．

変数名	コード	意味
Gakureki	E09504	最終学歴が大学・大学院卒の者の割合 (%)
Shushou	A05203	合計特殊出生率

3.2　分析

- 回帰分析（線形回帰モデル）

$$y = \beta_0 + \beta_1 x + \epsilon$$

- モデル式　`Shushou ~ Gakureki`
- 説明変数 $x =$ `Gakureki`，目的変数 $y =$ `Shushou`，誤差 ϵ，回帰係数 β_0, β_1.
- R の `lm()` で計算できる

3.3　プログラム

dat0001 に線形回帰モデルを当てはめ，学歴と出生率の関係を調べる．データの散布図に回帰直線を示す.

```
# run0001.R
# 回帰分析 (dat0001)
dat0001 <- read.table("dat0001.txt") # データの読み込み
print(dat0001) # 表示（対話的な実行時は自動的に print される）
plot(dat0001,pch=16) # 散布図
dat0001.lm <- lm(Shushou ~ Gakureki,dat0001) # 回帰分析の実行
print(summary(dat0001.lm)) # 結果の表示
abline(dat0001.lm,col=2,lty=2) # 回帰直線を引く
dev.copy2eps(file="dat0001-lm1.eps")
```

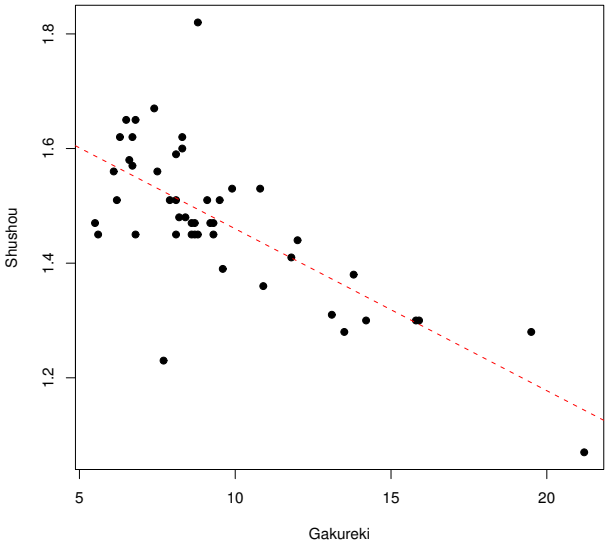
散布図に都道府県名を用いる.

```
# run0002.R
# テキストのプロット (dat0001)
plot(dat0001,type="n") # 枠だけ描く "n"は"no"の意味
text(dat0001,rownames(dat0001)) # 都道府県名
abline(dat0001.lm,col=2,lty=2) # 回帰直線
dev.copy2eps(file="dat0001-lm2.eps")
```

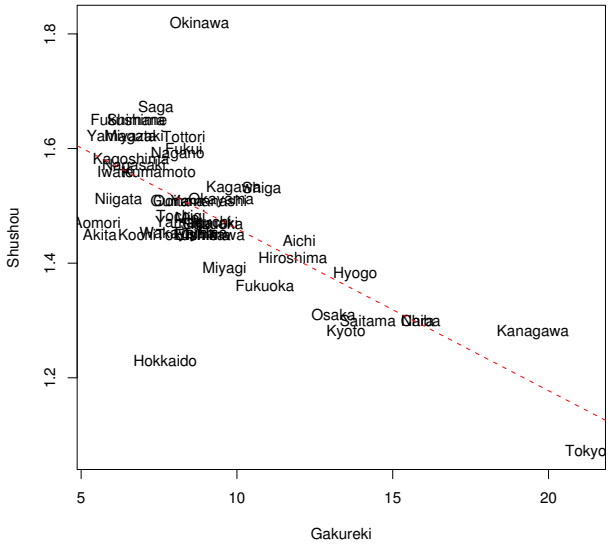
3.4 セッション

```
> source("run0001.R")
          Gakureki Shushou
Hokkaido       7.7    1.23
Aomori         5.5    1.47
Iwate          6.1    1.56
Miyagi         9.6    1.39
Akita          5.6    1.45
...中略...
Nagasaki       6.7    1.57
Kumamoto       7.5    1.56
Ooita          7.9    1.51
Miyazaki       6.7    1.62
Kagoshima      6.6    1.58
Okinawa        8.8    1.82

Call:
lm(formula = Shushou ~ Gakureki, data = dat0001)

Residuals:
      Min        1Q    Median        3Q       Max
-0.294968 -0.048132 -0.009319  0.045992  0.326105

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.742483   0.039973  43.592  < 2e-16 ***
Gakureki    -0.028249   0.003946  -7.158 5.94e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09205 on 45 degrees of freedom
Multiple R-Squared: 0.5324,Adjusted R-squared: 0.522
F-statistic: 51.24 on 1 and 45 DF,  p-value: 5.943e-09

> source("run0002.R")
```

3.5　結果



dat0001-lm1



dat0001-lm2

- モデル 出生率 $= \beta_0 + \beta_1 \times$ 学歴 $+ \epsilon$
- 回帰係数 $\hat{\beta}_0 = 1.74$, $\hat{\beta}_1 = -0.028$
- 回帰係数の標準誤差 $\hat{\sigma}_0 = 0.04$, $\hat{\sigma}_1 = 0.004$
- 回帰係数の確率値 $p_0 = 2 \times 10^{-16}$, $p_1 = 6 \times 10^{-9}$
- まとめ：　高学歴者の多い都道府県ほど出生率が低下する．大卒率が 10 ポイント増えると，出生率が 0.28 下がる．
- 注意：　これは都道府県の特徴を議論しているのであり，個人（又は世帯）における学歴と出生率の関係を議論しているのではない．いずれにしても，学歴と出生率の因果関係を示唆するとは限らない．

3.6　ニュース記事は注意して読む

２００４年９月２４日　社会ニュース

＜肺がん発生率＞幹線道路近くの住人で高く　胃がんも

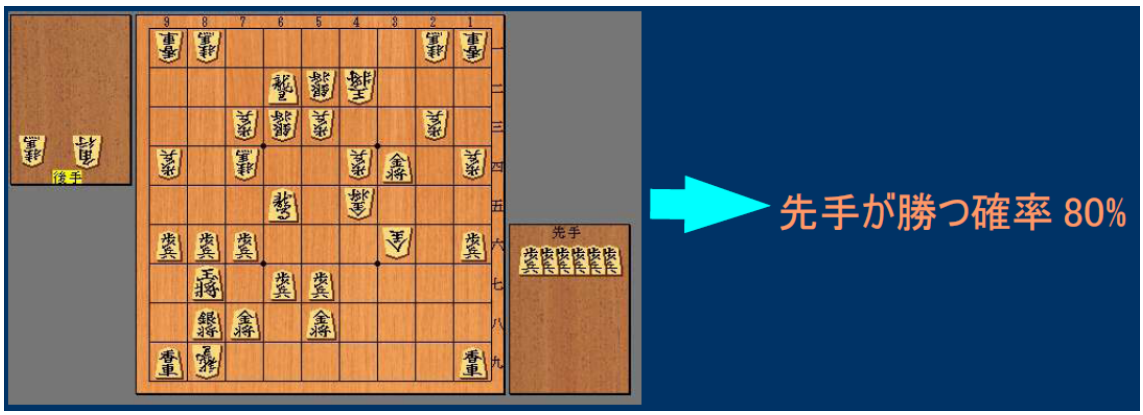
　幹線道路から５０メートル以内に住んでいる人は肺がんや胃がんになるリスクが高いことが、千葉県がんセンター研究局疫学研究部の三上春夫部長らの調査で分かった。男性の肺がんで１．７６倍、男女の胃がんで１．６８倍、それぞれ発生率が高くなっているという。２９日から福岡市で開かれる日本癌（がん）学会で発表する。　三上部長らは９０～９４年に同県内のある市で胃、大腸、肝、子宮、乳房のがんと診断された人のうち、１２時間の交通量が５０００台以上の幹線道路から５００メートル以内に住む５２８人について、幹線道路からの距離を精密に計測した。　続いて、当時の国勢調査に基づいた人口と実際の患者数から、５００メートル以内に住む人のがん発生率を割り出した。これをもとに５０メートル以内の発生数を予測し、実際の患者数と比べた。　この結果、予測発生数と実際の患者数は、男性の肺がんで９．６

４人と１７人、男性の胃がんで２２．０１人と３７人、女性の胃がんで１２．５４人と２１人だった。幹線道路から５０メートル以内に住む人はより遠くの住民よりも、発生率が男性の肺がんで１．７６倍、男女の胃がんで１．６８倍高いことになる。　他のがんでは、女性の肺がん２．００倍、男性の大腸がん１．３２倍、女性の大腸がん１．６２倍、男性の肝がん１．４６倍、女性の肝がん１．１９倍、乳がん０．８７倍、子宮がん１．０４倍――との結果だったが、患者数が少ないなどで統計的に意味のある数字にならなかった。　三上部長は「５０メートル以内に住むがん患者の年齢は全県平均より若く、交通量の多い幹線道路特有の事情があると考えられる。自動車の排ガスに含まれる有害成分が関与しているとみられるが、胃がんでもリスクが高くなっているので、単純に吸入だけの影響ではないようだ」と話している。【吉川学】（毎日新聞） - 9 月 24 日 3 時 5 分更新

- 幹線道路に近いと空気が悪いので肺がんになりやすい？ (因果関係？)
- 幹線道路に近いと騒音などストレスが多く胃がんになりやすい？
- 幹線道路の近くに住む人はどういう人？ (傾向に関連？)

3.7　卒論紹介 (I)

2004 年度学士論文
統計的学習を用いたゲーム勝敗予測とコンピュータ将棋への応用
情報科学科 下平研究室 谷口智也



- 勝敗予測関数の構成
  - ロジスティック回帰 (y は先手勝ち=1，負け=0)

$$\log \frac{P(Y=1|x)}{P(Y=0|x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

  - 変数選択 (x は盤面評価の特徴量を 369 変数)
- コンピュータ将棋への応用

4　多変量の関係を調べる（その１）

4.1　データ

dat0002: 日本の 47 都道府県について，以下の表にある１０変量の値．サイズ 47×10 の実数行列．後述する X2000 データセットの一部であり,「コード」は X2000 における変数のコード.(「65Sai」のように数字から始まる文字列をデータセットの変数名に用いることは避けたほうが良い．以下に見るように，多少面倒がおこる．)

変数名	コード	意味
Zouka	A05201	自然増加率 (%)
Ninzu	A06102	一般世帯の平均人員 (人:person)
Kaku	A06202	核家族世帯割合 (%)
Tomo	F01503	共働き世帯割合 (%)
Tandoku	A06205	単独世帯割合 (%)
65Sai	A06301	65 歳以上の親族のいる世帯割合 (%)
Kfufu	A06302	高齢夫婦のみの世帯の割合 (%)
Ktan	A06304	高齢単身世帯の割合 (%)
Konin	A06601	婚姻率（人口千人当たり）
Rikon	A06602	離婚率（人口千人当たり）

4.2　分析

- 主成分分析

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{10} x_{10} + \epsilon$$

- モデル式（新たな「合成変量」が左辺に対応する）

```
~ Zouka + Ninzu + Kaku + Tomo + Tandoku + `65Sai` +
   Kfufu + Ktan + Konin + Rikon
```

  または `~ .`．(最後の dot はすべての変数を意味する）．もしくはモデル式を省略しても良い.

- y は合成変量であり，変数の特徴を良く表すものが自動的に生成される．$x_1, \ldots, x_{10}$ はデータセットの１０変量.

- R の `princomp()` で計算できる

4.3　プログラム

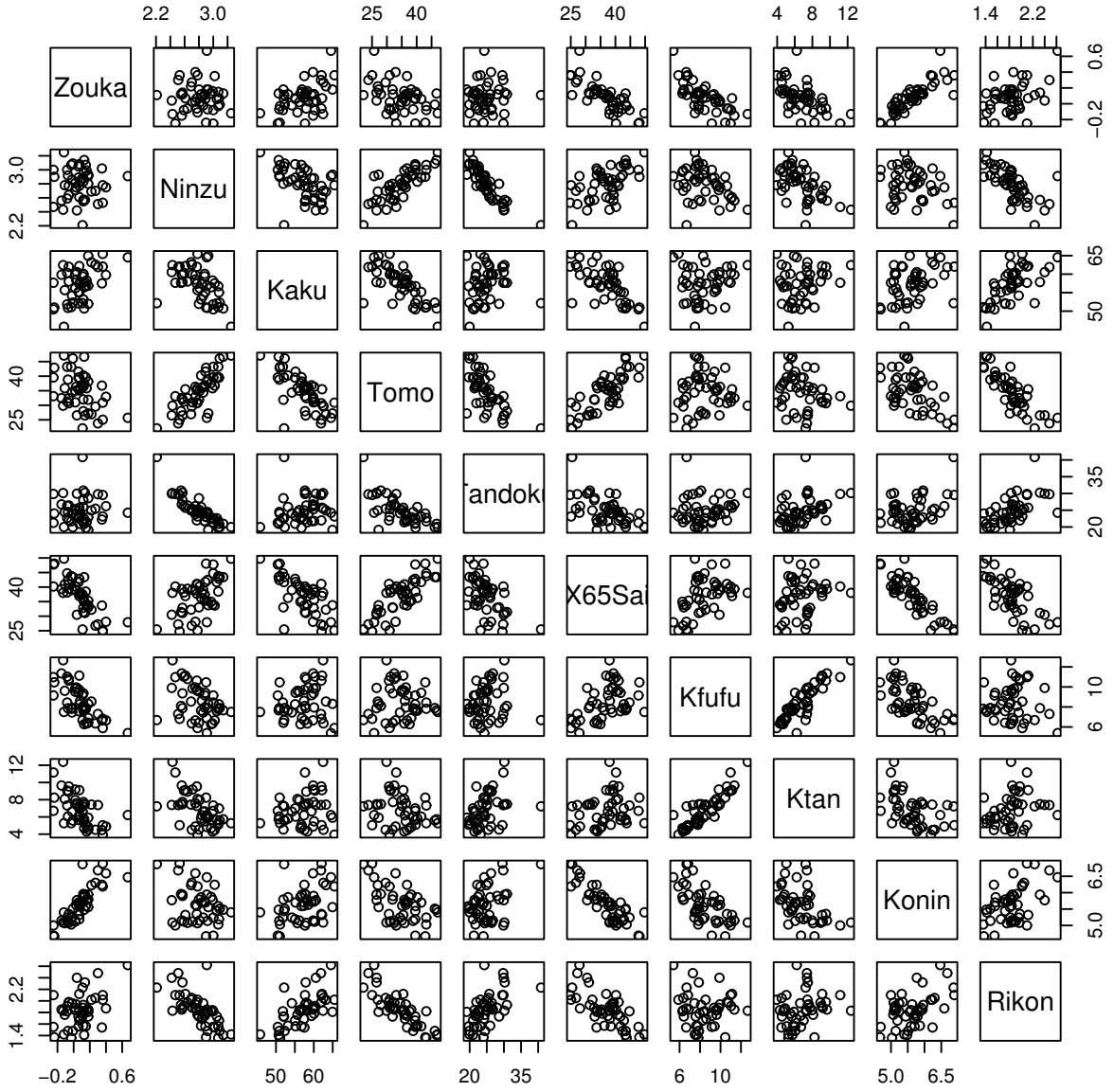
dat0002 の１０変量のすべてのペアの散布図をプロット．dat0002 に主成分分析を適用し，成分の分散のプロットと，第１，第２主成分のバイプロットを表示．`princomp` においてモデル式を省略してあるが，`princomp(~ .,data0002,cor=T)` などとしても同じ．オプション `cor=T` はあらかじめデータセットの変数の分散を１にそろえる標準化を実行することを意味する.

```
# run0003.R
# 主成分分析 (dat0002)
dat0002 <- read.table("dat0002.txt") # データの読み込み
print(dim(dat0002)) # データ行列のサイズを表示
print(dat0002[1:3,]) # データの要素を最初の３個だけ表示
pairs(dat0002) # 変量のすべてのペア毎に散布図を描く
dev.copy2eps(file="dat0002-sp.eps")
dat0002.pca <- princomp(dat0002,cor=T) # 主成分分析の実行
print(dat0002.pca)
plot(dat0002.pca)
dev.copy2eps(file="dat0002-pc1.eps")
biplot(dat0002.pca) # バイプロット
dev.copy2eps(file="dat0002-pc2.eps")
```

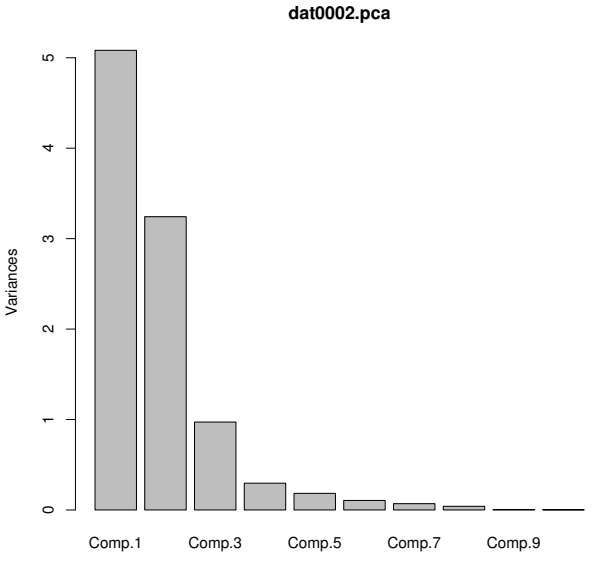
4.4 セッション

```
> source("run0003.R")
[1] 47 10
         Zouka Ninzu  Kaku  Tomo Tandoku X65Sai Kfufu Ktan Konin Rikon
Hokkaido  0.04  2.42 60.54 26.54   29.95  30.50  9.90 7.39  5.77  2.40
Aomori   -0.02  2.86 54.20 34.38   24.08  38.99  7.45 6.61  5.24  1.96
Iwate    -0.07  2.92 50.87 38.82   24.47  42.42  7.87 6.05  5.14  1.48
Call:
princomp(x = dat0002, cor = T)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
2.25433141 1.80067886 0.98597313 0.54426153 0.42863015 0.32479623 0.26359482
    Comp.8     Comp.9    Comp.10
0.20200083 0.06844848 0.05471515

 10  variables and  47 observations.
```

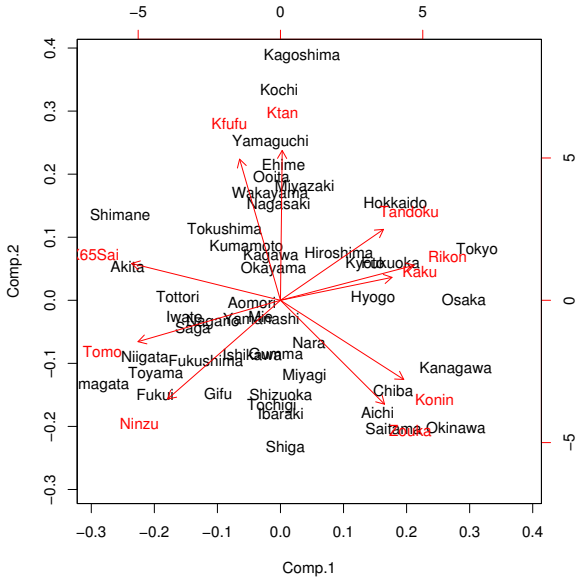
4.5 結果



dat0002-sp



dat0002-pc1

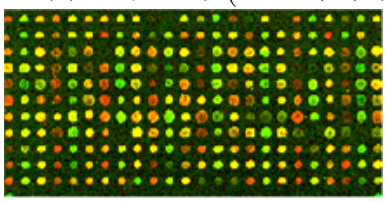


dat0002-pc2

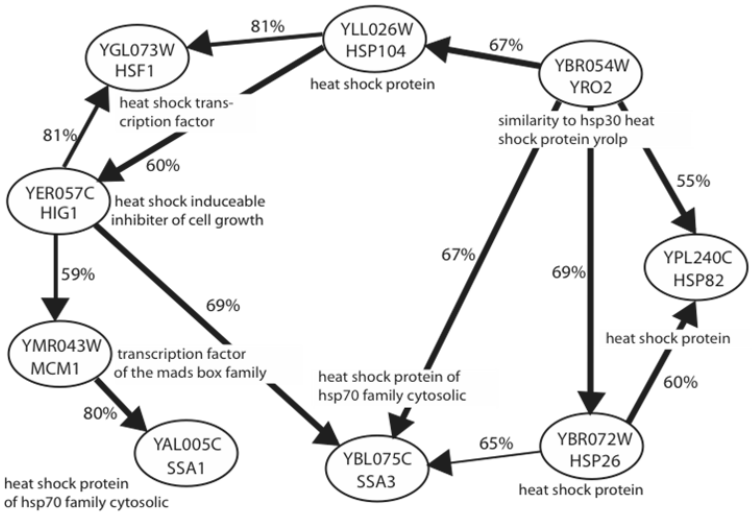
4.6 卒論紹介 (II)

2003 年度学士論文
DNA マイクロアレイデータに基づく遺伝子ネットワーク推定
情報科学科 下平研究室 上村健

- マイクロアレイ (DNA チップ) で遺伝子の発現レベルを測定する



- 遺伝子ネットワーク



- 遺伝子機能の解明，薬剤開発

5　多変量の関係を調べる（その２）

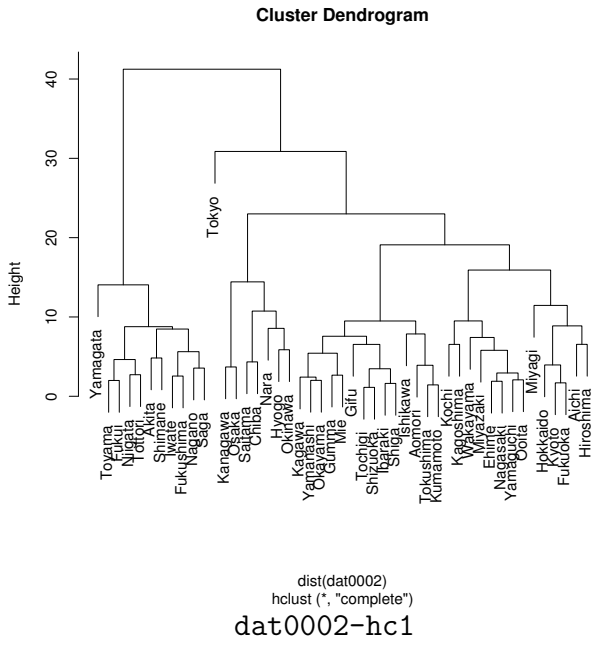
5.1　分析

- 階層的クラスタリング分析を前節の１０変量データセットに適用し，要素（都道府県）の関係や，変数間の関係を調べる．
- 要素間の「距離」行列を生成する関数 `dist()`
- 階層的クラスタリングを実行する関数 `hclust()`
- 結果の表示 `plot()`

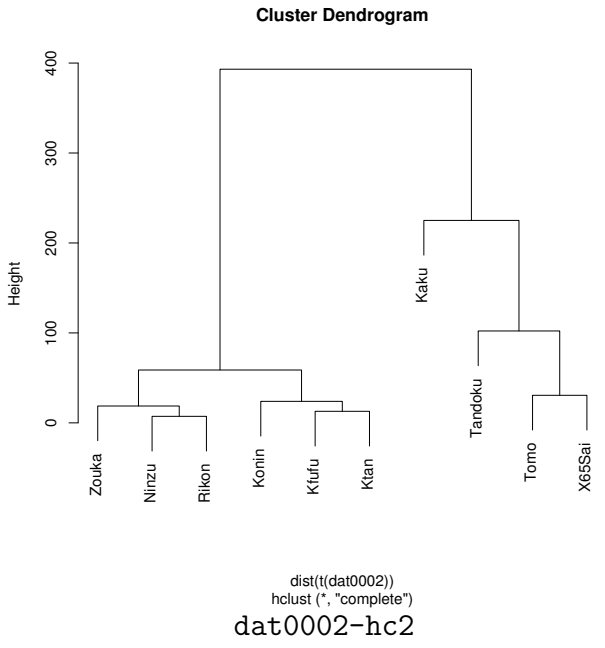
5.2　プログラム

```
# run0005.R
# クラスタ分析 (dat0002)
dat0002 <- read.table("dat0002.txt")
dat0002.hc <- hclust(dist(dat0002)) # クラスタ分析の実行
plot(dat0002.hc) # 結果のプロット
dev.copy2eps(file="dat0002-hc1.eps")
hc2 <- hclust(dist(t(dat0002))) # データ行列を転置してからクラスタ分析
plot(hc2) # 結果のプロット
dev.copy2eps(file="dat0002-hc2.eps")
```

5.3　結果



dat0002-hc1



dat0002-hc2

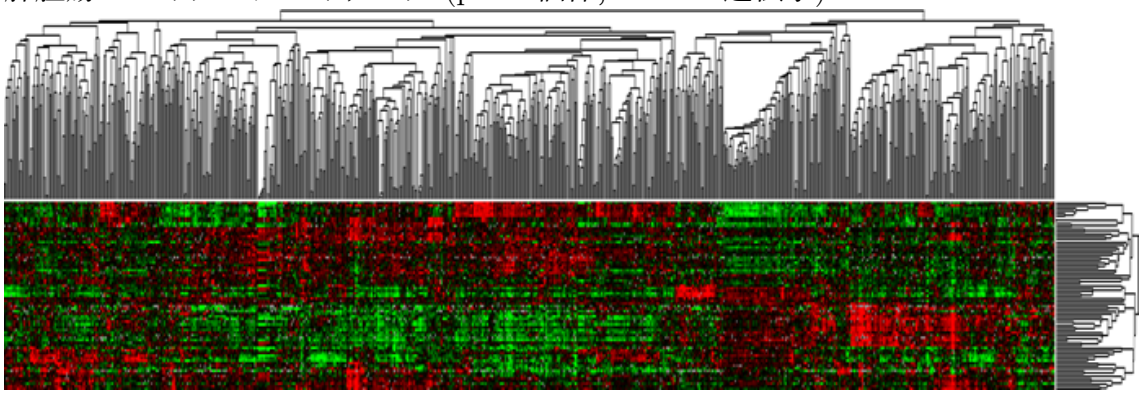
5.4 卒論紹介 (III)

2004 年度修士論文

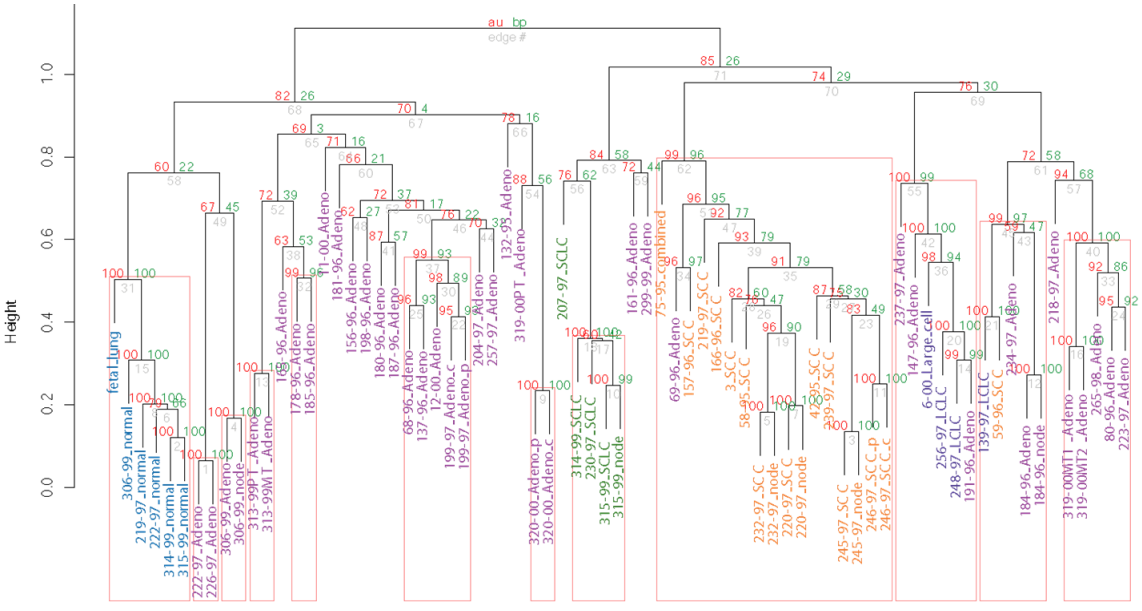
Assessing the uncertainty in hierarchical cluster analysis via multiscale bootstrap resampling (階層的クラスタリングの信頼性評価をマルチスケール・ブートストラップ法で行う)

数理計算科学専攻 下平研究室 鈴木了太

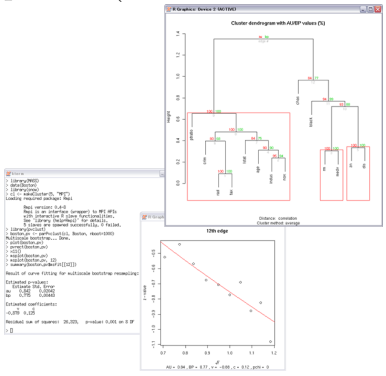
- 肺腫瘍のマイクロアレイデータ (p=73 個体，n=916 遺伝子)



- ７３腫瘍のクラスタリング



- pvclust (R の公式ライブラリに登録済み)



6 X2000データセット

6.1 データ

- データは総務庁統計局統計センターが公開している社会・人口統計体系である
  http://www.stat.go.jp/data/ssds/index.htm
  http://www.stat.go.jp/data/ssds/9.htm
- 総務庁が公開しているデータはExcel形式であるが，これを加工してテキストファイルにしたものを講義で用いる．
- 47都道府県の様々な調査項目 (2000年度の1182項目) を表にしたテキストファイルがX2000data.txtである．つまりサイズ 47×1182 の実数行列．各項目はA01101のようなコードによって表される．
- 補足情報のテキストファイルがX2000item.txtである．1182項目について，変数の意味 (Imi)，単位 (Tani)，全国値 (Zenkoku)，分類 (Bunrui) が示されている．つまりサイズ 1182×4 の行列である．ただし，要素は文字列や実数である．
- テキストファイルX2000code.txtに変数の「コード」,「意味（単位)」をまとめてある．

6.2 セッション

X2000.dataにデータセットを読み込み，県名や項目コードを確認.

```
> X2000.data <- read.table("X2000data.txt")
> dim(X2000.data)
[1]   47 1182
> rownames(X2000.data)[1:5]
[1] "Hokkaido" "Aomori"   "Iwate"    "Miyagi"   "Akita"
> colnames(X2000.data)[1:5]
[1] "A01101"   "A01601"   "A0160101" "A0160102" "A0160103"
> names(X2000.data)[1:5]
[1] "A01101"   "A01601"   "A0160101" "A0160102" "A0160103"
> X2000.data[,"E09504"]
 [1]  7.7  5.5  6.1  9.6  5.6  6.3  6.5  9.3  8.2  8.1 14.2 15.9 21.2 19.5  6.2
[16]  8.8  9.3  8.3  9.1  8.1  8.7  9.2 12.0  8.4 10.8 13.5 13.1 13.8 15.8  8.1
[31]  8.3  6.8  9.5 11.8  8.6  8.6  9.9  8.7  6.8 10.9  7.4  6.7  7.5  7.9  6.7
[46]  6.6  8.8
```

一部を取り出す（データセットdat0001）.

```
> x <- X2000.data[,c("E09504","A05203")]
> plot(x)
> names(x) <- c("Gakureki","Shushou")
> plot(x)
```

```
> fit <- lm(Shushou ~ Gakureki,x)
> abline(fit)
```

項目の詳細を確認

```
> X2000.item <- read.table("X2000item.txt")
> dim(X2000.item)
[1] 1182    4
> names(X2000.item)
[1] "Imi"     "Tani"    "Zenkoku" "Bunrui"
> X2000.item[c("E09504","A05203"),c("Imi","Tani")]
                                        Imi Tani
E09504 最終学歴が大学・大学院卒の者の割合（%）
A05203                          合計特殊出生率
> X2000.item[c("E09504","A05203"),"Zenkoku"]
[1] 11.90  1.36
> X2000.item[c("E09504","A05203"),"Bunrui"]
[1] E. 教育 7) 教育普及度       A. 人口・世帯 5) 人口動態
153 Levels: A. 人口・世帯 1) 人口の規模・構造 ... M. 生活時間 3) 3 次活動の種類別平均時間
```

7 課題

7.1 サンプル：課題 1-0

dat0001 の回帰分析で説明変数 x と目的変数 y を交換して，モデル式 Gakureki ~ Shushou を適用せよ．

7.2 課題 1-1

dat0002 の１０変量から自由に２変量を選び，線形回帰分析（単回帰）を行え．分析に用いたプログラム，セッション，結果（グラフ，回帰係数等，まとめ）を示すこと．

7.3 課題 1-2*

X2000 から自由に２変量を選び，線形回帰分析を行え．変数の「コード」と「意味」を明記すること．

7.4 課題 1-0 の略解

7.4.1 プログラム

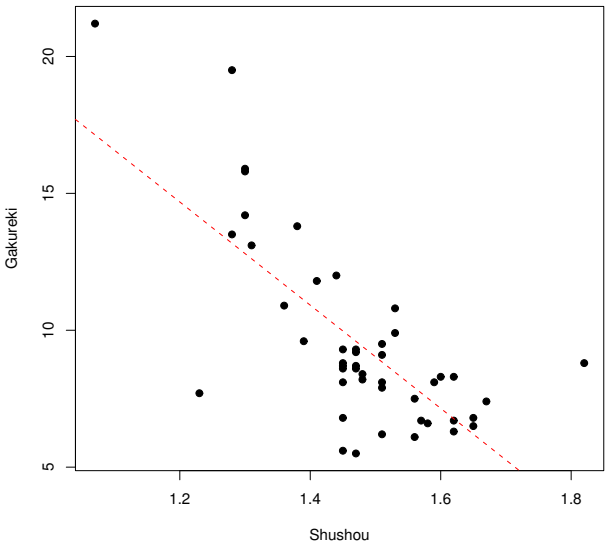
```
# run0004.R
```

```
# 回帰分析 (dat0001)  x と y の交換
dat0001 <- read.table("dat0001.txt")
plot(Gakureki ~ Shushou,dat0001,pch=16) # 横軸=Shushou, 縦軸=Gakureki
fit <- lm(Gakureki ~ Shushou,dat0001) # 回帰分析の実行
print(summary(fit))
abline(fit,col=2,lty=2) # 回帰直線
dev.copy2eps(file="dat0001-lm3.eps")
plot(Gakureki ~ Shushou, dat0001,type="n") # 枠だけ描く
text(dat0001[,"Shushou"],dat0001[,"Gakureki"],rownames(dat0001)) # 都道府県名
abline(fit,col=2,lty=2) # 回帰直線
dev.copy2eps(file="dat0001-lm4.eps")
```

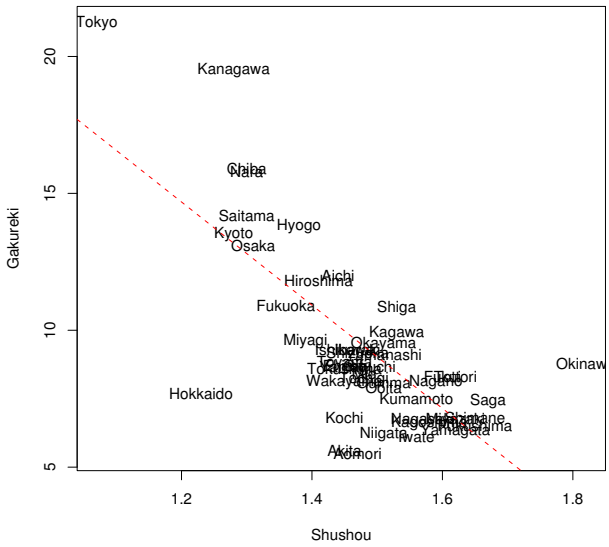
7.4.2 セッション

```
> source("run0004.R")

Call:
lm(formula = Gakureki ~ Shushou, data = dat0001)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4199 -1.0927 -0.3966  1.2765  6.3225

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   37.302      3.894   9.580 1.96e-12 ***
Shushou      -18.847      2.633  -7.158 5.94e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.378 on 45 degrees of freedom
Multiple R-Squared: 0.5324,Adjusted R-squared: 0.522
F-statistic: 51.24 on 1 and 45 DF,  p-value: 5.943e-09
```

7.4.3 結果



dat0001-lm3



dat0001-lm4

- モデル 学歴 $= \beta_0 + \beta_1 \times$ 出生率 $+ \epsilon$
- 回帰係数 $\hat{\beta}_0 = 37.3$, $\hat{\beta}_1 = -18.8$
- 回帰係数の標準誤差 $\hat{\sigma}_0 = 3.9$, $\hat{\sigma}_1 = 2.6$
- 回帰係数の確率値 $p_0 = 2 \times 10^{-12}$, $p_1 = 6 \times 10^{-9}$
- まとめ： 出生率が高い地域ほど高学歴者が少ない．出生率が 1 増えると大卒率が約 19 ポイント下がる．