第２講

目次

第２講

1.多変量データの扱い

本章の Point	多変量のデータ分析手法として、重回帰分析と主成分分析、クラスター分析の理解を深めます。また、実際のデータを用いた分析を紹介し、統計データでの分析の実習ができるようなります。
利用 Data	✧　理科年表「世界の気象データ」 ✧　総務省「全国消費実態調査」

（1）重回帰分析

本項における参考情報	✧　日本統計協会　統計学Ⅲ「多変量データ解析法　オフィシャル　スタディノート」 ✧　総務書統計局統計研修所　「政策立案と統計に係る応用的研修テキスト」

複数の変数の関係が知りたい

- 第 1 講で扱った単回帰分析では、データの散らばりの中から、（１つの）説明変数と被説明変数（目的変数)の関係を表す直線を、各観測値における被説明変数の値と直線上の理論値との間のずれの二乗の総和が小さくなるような場所で描画し、その推定線の傾き、切片から推定式を導きました[1]。（第 1 講「単回帰分析」参照）

- 重回帰分析とは、単回帰分析の説明変数を２つ以上にしたものです。被説明変数（課題意識を持っている事象を占める指標）と複数の説明変数（要因と考えられる要素）の量的な関係（相関関係）を把握できます。

$$\text{重回帰分析}\quad y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + u_i$$

[1] この手法を「最小二乗法」といい、重回帰分析のほか、直線でない回帰にも応用できます。

＜重回帰分析とは＞

重回帰分析とは、被説明変数 y （結果）と 2 つ以上の説明変数 x （原因）との関係を測定する回帰分析のことです。単回帰分析の説明変数を 2 つ以上に拡張したものを重回帰分析といい、多変量解析の一つの手法です。特に、説明変数、被説明変数が量的なデータである場合に用います。

「原因」、「結果」としましたが、重回帰分析では因果関係を厳密に見分けることはできません。ただし、数式にデータを当てはめる際に「恐らくこれは原因であろう」、「恐らくこれは結果であろう」と考えられるものをそれぞれ説明変数、被説明変数とすることで、説明変数（原因）で被説明変数（結果）をどの程度説明できるかを示すことができます。

＜変数の選択＞

重回帰分析で用いる説明変数同士には、相関関係がなさそうな（相関係数の小さい）ものを選択することに注意する必要があります。相関関係の強い説明変数を選択すると、回帰係数の分散が大きくなり推定ができない、又は、推定された効果が過大に推計されてしまう多重共線性の問題が発生します。多重共線性とは、「Multicollinearity マルチコリニアリティ（単にマルチコ）」とも呼ばれ、説明変数間で、相関係数が大きい組み合わせがあることです。多重共線性が発生している場合、説明変数の回帰係数は過大に推計されます。

＜重回帰分析の実施＞

重回帰分析の数学的・理論的背景の理解は、線形代数の細かい知識が必要となり、難しいのですが、表計算ソフト等に関数等が実装されているので、概念を理解すればすぐに利用することができます。

上図のような回帰式は、EXCEL の「分析ツール」から「回帰分析」を選択することで分析ができます。分析の関心となる出来事の結果に関する変数を y （被説明変数）、分析の関心となる出来事の原因と考えられる変数（複数選択が可能）を x （説明変数）として、データを当てはめていきます。また、本項の末尾に Python、R での推定の実施例を記載しておりますので、そちらもご参照ください。

重回帰分析結果を確認し、その内容を理解するために必要な各変数は、次のように整理します。

重回帰分析における変数

名称	記号	数値の持つ意味
被説明変数	y	分析の関心となる出来事の結果となる変数です（従属変数、目的変数とも言います）。
説明変数	x	分析の関心となる出来事の原因と考えられる変数（独立変数とも言います）で、重回帰なのでベクトル（多変量変数）となります。
定数項	a	説明変数からの影響が一切ない場合の、被説明変数の理論的な値となります。ダミー変数等を入れると意味合いが変わるので注意が必要です。詳細は他の参考書を参照ください。
回帰係数	β	説明変数が 1 単位だけ増えると、被説明変数 Y がどれだけ増えるかを表します。
誤差項	u	被説明変数の変化が、説明変数の変化だけでは説明できない、理論と結果との間に生じるズレを表す部分です。（残差とも言う）
決定係数	R^2 (R2)	決定係数はモデルの説明力を表す値です。ここで説明力とは被説明変数の散らばりのどれほどが説明変数で説明できるかを表すもので、1 に近づくほど被説明変数により説明できる割合が高いといえます。
自由度調整済み決定係数	$AdjR^2$ （補正 R2）	決定係数はその数式の特性上、p 数が増えると自動的に決定係数も多くなっていく。これを自由度に応じて修正したものが自由度調整済み決定係数です。判断基準は決定係数と同じです。

分析の目的とその解釈

重回帰分析は、ソフトウェアで簡単に求めることができますが、その出力結果が何を示しているかを解釈するには一筋縄ではいきません。回帰分析を行う目的によって、その出力結果の解釈は異なります。

重回帰分析の目的は予測と制御に分けられます。

分析の目的	何を知りたい	結果の見方
予測	新しく与えられた説明変数$(x_1, \cdots, x_n)$の値から被説明変数yの値を知ろうとすること	良い予測モデルを立てるために$(\beta_1, \cdots, \beta_n)$の値を用いる。回帰係数の解釈は本質的ではないため、回帰係数$(\beta_1, \cdots, \beta_n)$の値の解釈はしない。

制御	回帰係数$(\beta_1, \cdots, \beta_n)$の値を見て、被説明変数$y$の最適値を与える説明変数$(x_1, \cdots, x_n)$を定めること	制御に用いる変数を、$(\beta_1, \cdots, \beta_n)$を見て選択する。例えば$\beta_i$の値がほとんど 0 に近い変数は$y$への影響力が小さい、といった解釈がなされる。

＜変数を増やして重回帰分析を行う＞

実際のデータ分析では、1 つの現象を 1 つの出来事で説明することは困難です。1 つの現象に複数の理由がある場合は、説明変数を増やして、重回帰分析を行います。

また、むやみやたらに説明変数を増やせばいいというわけでもありません。重回帰分析で変数選択を行う際に気を付けるべきことは、説明変数同士の相関関係がないということです。説明変数同士に相関がみられると前述した多重共線性問題が生じて、推定結果が正しくデータ間の関係を示さない結果となります。モデル構築の際に説明変数の組み合わせを増やす場合は説明変数同士の関係についても留意します。

＜説明変数の取捨選択＞

重回帰分析の説明変数を増やした後は、必要な変数と不要な変数を選ぶ「変数選択」が必要となります。変数選択とは、説明変数あるいはモデルの候補の中から、被説明変数yを最もよく説明する変数の組を同定することです。ここでも、前述した分析の目的によって、変数選択の方法が異なってくることに注意が必要です。すなわち、変数の性格によって選択すべき、又はすべきでない変数が決まります。

しかし、どの変数を選択するべきか否かを考える前に、そもそも「変数選択が必要であるかどうか」も問う必要があります。変数選択が不要な場合とは、以下のような場合です。

変数選択が不要な場合

- **説明変数が決まっている場合**
  - 与えられた変数群から被説明変数の値を予測すればよいので変数選択は不要です。
- **分析の目的が予測の場合**
  - 使える変数はなるべく使う方が良く、不要な変数があってもあまり問題はありません。ただしなんでも変数として足せばいいというわけではありません。
- **被説明変数の最適値を求める場合**
  - サンプルサイズがあまりにも少なすぎる場合ではない限り、変数選択は不要です。

変数選択が必要な場合

- **分析の初期段階での変数を選択する場合**
  - 候補となる変数群から必要なものを選択します。
- **予備調査の結果を用いて本調査を行う場合**
  - 本調査で理論上あるいは実質上必要なもののみ選択します。多くの項目を調査することはコスト増加につながるためです。
- **制御すべき変数を同定する場合**
  - 多くの変数の制御は困難であるので、なるべく少数の変数に絞る必要があります。

実際のデータを用いて、重回帰分析を行ってみましょう。「高校からの統計・データサイエンス活用」に世界各国の都市において、平均気温と緯度、標高との関係を計算で導く方法として、重回帰分析が紹介されています。

理科年表から、以下のような各国の都市の平均気温と緯度、標高を利用します。

番号	地点	国または領域	緯度	経度	高度	気温	降水量	相対湿度	気圧
1	OSLO/GARDERMOEN	ノルウェー	60 12 N	11 04 E	202	4.8	849.7	79	1020.4
2	STOCKHOLM/BROMMA	スウェーデン	59 21 N	17 53 E	15	6.7	535.8	76	1012
3	HELSINKI-VANTAA	フィンランド	60 19 N	24 58 E	51	5.3	678.6	82	1020.3
4	HEATHROW	イギリス	51 28 N	00 27 W	24	11.8	640.3		1020.6
5	DUBLIN AIRPORT	アイルランド	53 26 N	06 15 W	68	9.8	775.2		1014.3
6	REYKJAVIK	アイスランド	64 08 N	21 54 W	54	4.7	847.1	78.7	1007.9
7	NUUK (GODTHAAB)	グリーンランド	64 10 N	51 45 W	80	-1.4	611.6	81	1020.8
8	KOEBENHAVN/LANDBOHOEJSKOLEN	デンマーク	55 41 N	12 32 E	7	9.1	582.1		1016.8
9	DE BILT AWS	オランダ	52 05 N	05 10 E	1	10.1	828.5	82	1016.2
10	UCCLE	ベルギー	50 48 N	04 21 E	99	10.6	855.1	81.6	1012.7

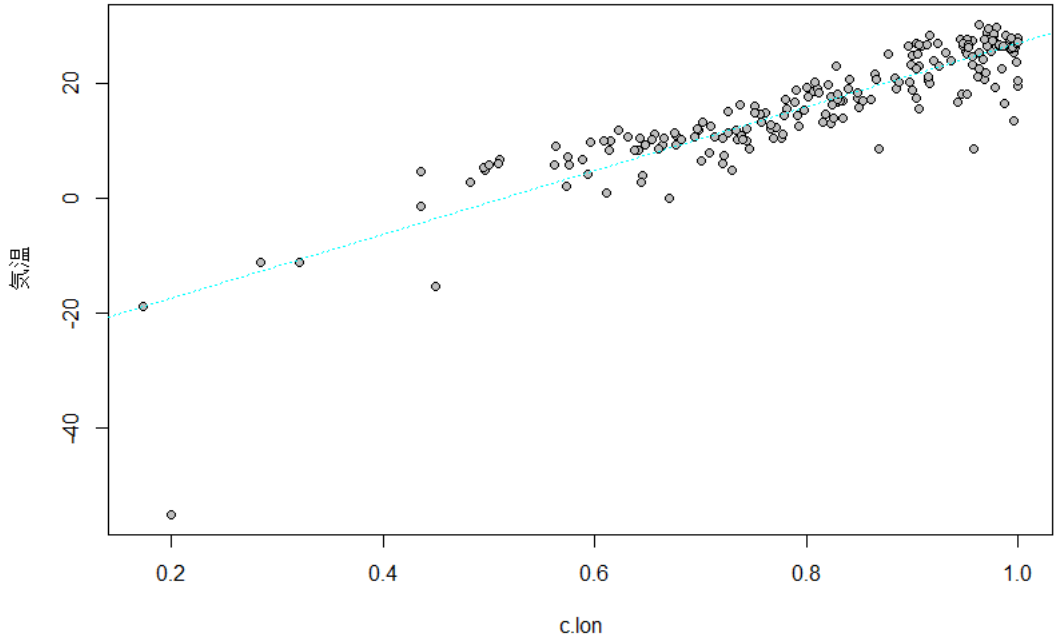
出典：国立天文台編「理科年表 2020」, 丸善出版 (2019).理科年表オフィシャルサイト(国立天文台・丸善出版).
National Astronomical Observatory of Japan, Chronological Scientific Tables, Maruzen, (2019).

平均気温(temp)を被説明変数、変換済みの緯度(c.lon)[2]を説明変数として、R プログラムで単回帰した結果表と散布図を以下に示します。（R および Python のコードは後掲）

```
Call:  lm(formula = 気温 ~ c.lon, data = dt1)

Residuals:      Min      1Q  Median      3Q     Max
              -37.772  -1.382   0.546   2.424   8.987

Coefficients:           Estimate Std. Error t value Pr(>|t|)
         (Intercept)  -28.571      1.560   -18.32  <2e-16 ***
           c.lon       55.657      1.868    29.79  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.488 on 212 degrees of freedom
  (26 observations deleted due to missingness)
Multiple R-squared:  0.8072,    Adjusted R-squared:  0.8063
F-statistic: 887.5 on 1 and 212 DF,  p-value: < 2.2e-16
```

主に確認が必要な t 値（t-value）と自由度調整済み決定係数（Adjusted R-squared）を赤の点線で囲みました。t 値は十分大きく、自由度調整済み決定係数も比較的大きく出ています。

[2] 緯度の例えば 60°12'N という表記は、北緯 60 度 12 分を示し、60 進数のためそのまま数値として計算するのには向きません。このため、これを、「cos（（度（°）＋（分（'）*1/60））*2*（円周率）/360*（南緯（S の場合）-1））」の形に変換します。



さらに、変数の追加を考えてみましょう。データセットにある他の変数は、標高（データセット上の項目名は「高度」）、降水量、相対湿度と気圧がありますが、この中で標高の次に温度に影響があると思われる項目が標高です。回帰分析の説明変数に標高（高度）を追加して分析した結果が以下のようになります。赤の点線部分が、主に確認が必要な t 値（t-value）と自由度調整済み決定係数（Adjusted R-squared）です。t 値は十分大きく、決定係数が 0.908 と上昇し、標高として追加した変数「高度」の t 値も有意で、モデルとしての精度が上がっていることが分かります。

```
Call:  lm(formula = 気温 ~ c.lon + 高度, data = dt1)
Residuals:       Min       1Q   Median       3Q      Max
              -21.1342  -1.5395   0.0741   1.7064   7.6163
Coefficients:             Estimate    Std. Error   t value  Pr(>|t|)
          (Intercept) -2.701e+01  1.079e+00  -25.02  <2e-16 ***
            c.lon       5.587e+01   1.287e+00   43.40  <2e-16 ***
            高度        -5.230e-03   3.407e-04  -15.35  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.092 on 211 degrees of freedom
  (26 observations deleted due to missingness)
Multiple R-squared:  0.9089,   Adjusted R-squared:  0.908
F-statistic:  1053 on 2 and 211 DF,  p-value: < 2.2e-16
```

このようにまずは単回帰分析を行いながら、データの背景やモデルの変動要因を理解し、適切に説明変数を加工、追加することでより現実に照らして妥当なモデルを作り出すことができます。

Python、R での重回帰分析

以下は、重回帰分析のグラフを作成するコードです。

実施内容	重回帰分析
利用 Data	世界の気象データ 理科年表 2019

Python コード（サンプル）

```
# 図やグラフを図示するためのライブラリをインポートする。
import matplotlib.pyplot as plt
from matplotlib.font_manager import FontProperties
# フォントの指定 C:¥WINDOWS¥Fonts¥YuGothR.ttc'部分はご利用のコンピュータにインストールされている日本語フォントファイルを指定してください。
fp = FontProperties(fname=r'C:¥WINDOWS¥Fonts¥YuGothR.ttc', size=14)

# 重回帰分析　必要なライブラリ
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import numpy as np

# データの読み込み
df = pd.read_csv("xxx.csv", index_col=0, encoding='shift_jis') # xxx 部分で csv を指定

# データ整形
df = df.dropna()
print(df.head())

# cos 計算を挟んだを説明変数、気温を被説明変数とした単回帰分析
x = df["ido_cos"]
X = sm.add_constant(x)
Y = df['kion']
model = sm.OLS(Y, X)
result = model.fit()

# 散布図を描く
plt.plot(x,Y,"o")
plt.ylim([-60,50])

#散布図に線形近似した直線を引く
plt.plot(x, np.poly1d(np.polyfit(x, Y, 1))(x), label='d=1')
plt.show()

#結果を出力
print("/////////////model1/////////////")
print(result.summary())
```

```
# cos 計算を挟んだ緯度と標高を説明変数、気温を被説明変数とした重回帰分析
a = df[["ido_cos","koudo_m"]]
A = sm.add_constant(a)
B = df['kion']
model = sm.OLS(B, A)
result = model.fit()

#結果を出力
print("/////////////model2/////////////")
print(result.summary())
```

※上記プログラムの緯度は、60°12'N といった形式の値を、「cos（（度（°）＋（分（'）*1/60））*2*（円周率）/360*（南緯（S の場合）-1））」で再計算して使用しています。

R コード（サンプル）

```
#データの読み込み
setwd("C:xxx") #xxx 部分でフォルダを指定
dt1 <- read.csv("世界の気象データ.csv", header=TRUE, as.is=TRUE)　　# 因子化抑制
tail (dt1)　# 内容確認
str(dt1)　# データ構造確認
dt1$気圧 <- as.numeric(dt1$気圧)　#　気圧データが文字属性なので、数値属性に修正
　　# 欠測があるので警告メッセージが出るが問題ない
attach(dt1)

# 緯度経度データを数量データとして扱えるように変換する
# まず半角ブランクで緯度経度データを分割
long <- strsplit(緯度, " ")
lati <- strsplit(経度, " ")
is.list(long)　　# long はリスト属性になる [1] TRUE
long2 <- t(matrix(unlist(long), nr=3))　　#　文字属性の行列
long3 <- round(as.integer(long2[,1]) + as.integer(long2[,2]) / 60, digits=2)
SN <- (long2[,3] == "N")　　# 北緯かどうか
lon <- long3 * SN + long3 * (SN-1)　　# 北緯はそのまま、南緯はマイナス値に

#　地点名でプロット
plot(lon, 気温, col="white")
par(new=T)
text(cbind(lon, 気温), 地点, cex=0.6)　　# 地点名でのプロット

#　緯度のコサインをとる
c.lon <- cos(2*pi*lon / 360)
cor(c.lon, 気温, use="complete.obs")　　# NA 除去
#[1] 0.898432　　# 相関高い
plot(c.lon, 気温, pch=21, bg="gray")　　# 散布図出力

#　cos(緯度)を説明変数、気温を目的変数とした単回帰分析
```

```
result1 <- lm(気温 ~ c.lon, data=dt1)
abline(result1, col="cyan", lty=3)                    # 散布図に水色点線で回帰線を追加する
summary(result1)

# 高度（標高）を追加して重回帰分析
result2 <- lm(気温 ~ c.lon+高度, data=dt1)
summary(result2)
```

※上記プログラムの緯度は、60°12'N といった形式の値を、「cos(（度（°）＋（分（'）*1/60））*2*（円周率）/360*（南緯（S の場合）-1））」で再計算して使用しています。

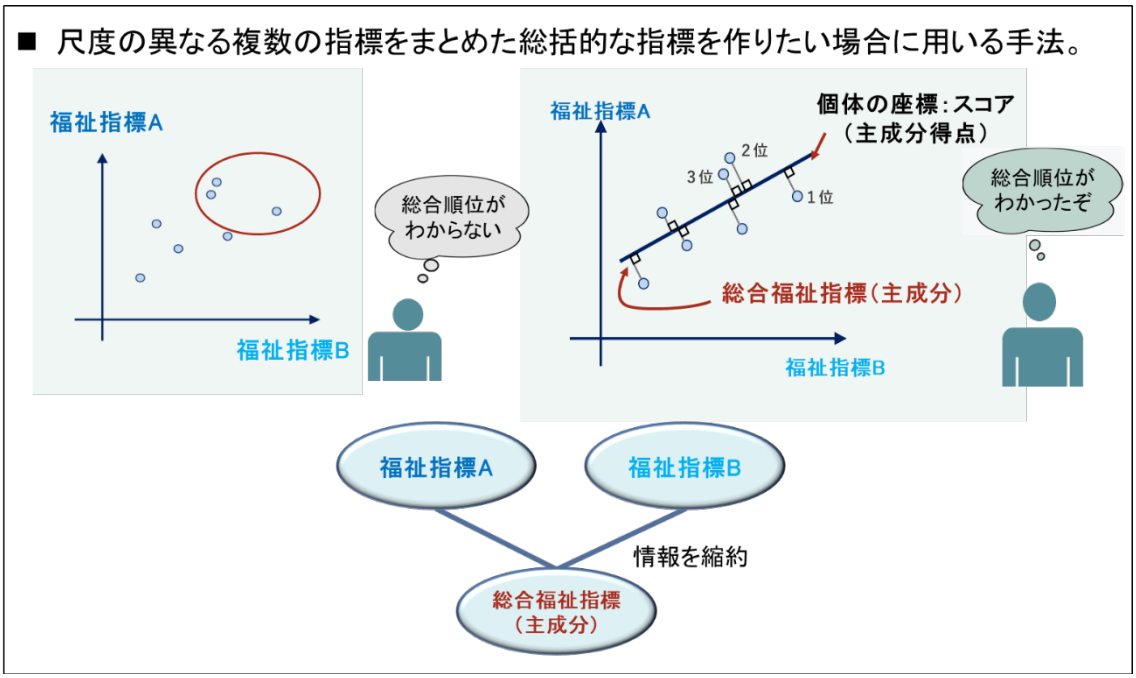
チャレンジ事項

世界の気象データを用いた分析方法を活用して、身近な課題での分析を行ってみましょう。e-Stat 等の公的統計 csv データを用いて重回帰分析を行い、得られた示唆について整理し考察してみましょう。また、説明変数を増やす過程で、「多重共線性」というものが生じることがあります。「多重共線性」とはどのような問題でしょうか。具体的な事例を Web で調べてみましょう。

（2）主成分分析

主成分分析の概略

本項における参考情報	✧ 日本統計協会　統計学Ⅲ「多変量データ解析法　オフィシャル　スタディノート」 ✧ 総務書統計局統計研修所 「政策立案と統計に係る応用的研修テキスト」 ✧ 静岡県「ふじのくに少子化突破戦略の羅針盤」 https://www.stat.go.jp/info/guide/rikatsuyou/pdf/sizuokaken_2017.pdf



＜主成分分析とは＞

(1)で解説した重回帰分析は、一つの被説明変数（目的変数）を複数の説明変数（独立変数）で説明する手法でした。説明変数同士は相関がないことが望ましいのですが、相関があるかないかは判断しにくいことが多いです。

これに対し、たくさんの変数からなるサンプルを対象とするとき、各データの持つ情報をできるだけ失うことなく、互いに相関のない、少ない変数にまとめる手法が主成分分析です[3]。主成分は、変数を定数倍したり変数同士を加減したり[4]して人工的に作り出される変数で、先に生成されたものと相関のないものの中で分散が最大になるように生成されます。新たに生成された変数を通して、データの中から有益な情報を取り出すことができます。第 1 主成分、第 2 主成分等の設定ができ、分析担当者の関心のある分析軸を設定することができます。一方で、分析担当者が各主成分にどのような情報が強く反応しているかを判断し、主成分の意味づけを行うため、分析担当者の興味関心をはっきりさせることや、要約しても違和感のない被説明変数群を設定することに留意します。例えば、福祉政策に関心がある場合に、被説明変数群は福祉政策に関係ある指標を選択する必要があります。仮に、福祉に全く関係のない変数を入れて主成分分析を行った場合、その分析結果を合理的に説明ができません。

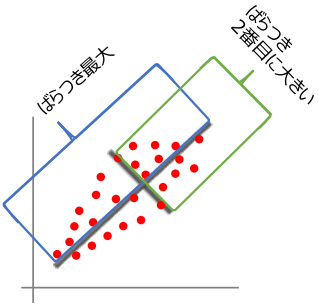
＜主成分分析の計算＞

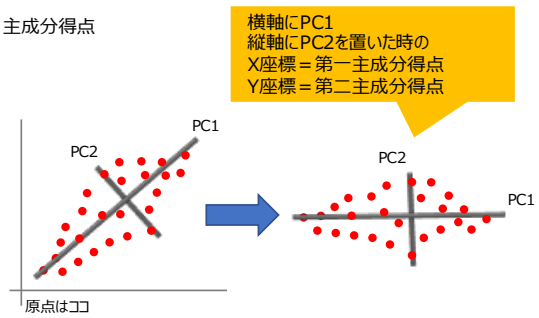
主成分分析を実施するには、福祉指標 A（以下、変数 A）、福祉指標 B（以下、変数 B）の主成分と主成分得点を計算する必要があります。主成分得点とは、各主成分軸を取った時の各データの

[3]主成分分析は数多い指標を比較的少数の指標にまとめることを目的としており、重回帰分析のような被説明変数（目的変数）は存在しません。

[4] このような操作を「線形変換」といいます。

座標に相当する値です。主成分軸は、分散が大きい順に直線を引くことで得られ、1 つのデータから複数の主成分軸を得ることができます。最も分散が大きい軸を第一主成分軸、2 番目に分散が大きい軸を第二主成分軸と呼びます。以下の図では、第一主成分軸は PC1、第二主成分軸は PC2 で表されています。一般的に主成分軸の数はデータの次元の数の増加に従い増加することが多いです。





主成分得点の計算方法は以下の通りとなります。

- 変数ごとに、平均値と標準偏差を用い、各データを標準化 [5] する。
- 各変数の相関行列、相関行列の固有値、固有ベクトルを求める [6]。
- 求めた最大固有値に対応する固有ベクトル、最大から 2 番目の固有値に対応する固有ベクトルをもとに、2 次元の座標を作り、点グラフをかく。
- 上記の作業で作った座標を使って標準化した変数を重みづけする。重みづけされた変数 A、B を加算して、主成分得点を作成する。

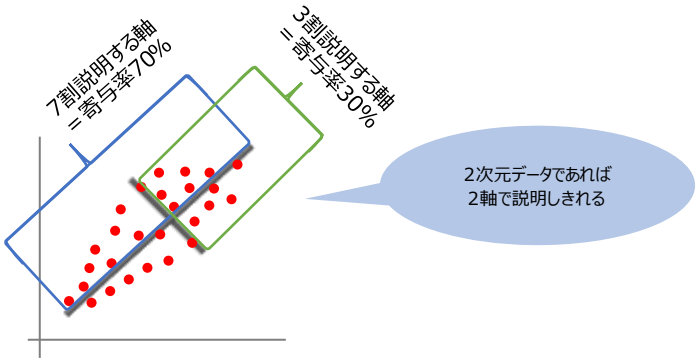
これにより、変数 A と変数 B の情報を要約した分析結果が得られます。分析結果の妥当性は累積寄与率 [7] を検討することで判断することが可能です。

データが 2 次元ならば、主成分軸を 2 つ設定するだけでデータの 100%が説明できます。一方で、データが 3 次元以上であれば、2 つの主成分軸では説明が不十分となる場合があります。しかし、次元数と同じだけ主成分軸を用意すると結果の解釈は困難を伴います。そのため、累積寄与率を確認しながら、主成分軸の数とモデルの妥当性を勘案して設定することが必要です。

[5] 標準化された A_i ＝ （A_i － 指標 A の平均値）/指標 A の標準偏差 （A_i はある市区町村 i の指標）

[6] 相関行列は変数同士の相関係数を並べた行列、固有値、固有ベクトルは行列 B に対し、$\mathbf{Bx} = \lambda\mathbf{x}$ となるベクトル x 及び単一の値（スカラー）λのことです。

[7] 累積寄与率は、設定した主成分の寄与率を加算することで算出することができます。（第 1 主成分の寄与率＝第 1 主成分の固有値÷設定した変数の個数×100）



上記のような一連の流れで情報を要約することを「多変量データの次元の縮約」といい、主成分分析は、多変量データの情報量を損なわずに情報を圧縮することが可能な手法であると言えます。

分析事例１：都道府県別食料品消費傾向

実際のデータを用いて主成分分析を実施します。今回は総務省統計局が調査している全国消費実態調査から、2014 年結果の第 13 表データ [https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200564&tstat=000001073908&cycle=0&tclass1=000001073965&tclass2=000001074840&tclass3=000001077457] を用いて各都道府県の食料品の消費に関する分析を行います。

＜使用した統計の特徴＞

統計名	全国消費実態調査（2019 年からは全国家計構造調査）
所管	総務省統計局
目的	全国消費実態調査は、国民生活の実態について、家計の収支及び貯蓄・負債、耐久消費財、住宅・宅地等の家計資産を総合的に調査し、全国及び地域別の世帯の消費・所得・資産に係る水準、構造、分布などを明らかにすることを目的とした調査である。また、全国消費実態調査は、昭和 34 年の第 1 回調査以来 5 年ごとに実施されており、平成 26 年全国消費実態調査はその 12 回目に当たる。
調査対象	全国の全ての世帯のうち、総務大臣の定める方法により選定された世帯を対象とし、二人以上の世帯と単身世帯とに分けて調査を実施した。 なお、次に掲げる世帯は、世帯としての収入と支出を正確に計ることが難しいこと等の理由から調査の対象から除外した。 （1）二人以上の世帯 a.料理飲食店又は旅館を営む併用住宅の世帯 b.下宿屋又は賄い付の同居人のいる世帯 c.住み込みの雇用者が 4 人以上いる世帯 d.外国人世帯 （2）単身世帯 a.二人以上の世帯の対象除外（a、b 及び d）に該当する者 b.学生の単身者

	c.15 歳未満の単身者 d.雇用者を同居させている単身者 e.社会施設及び矯正施設の入所者 f.病院及び療養所の入院者 g.自衛隊の営舎内居住者
調査事項	（1） 家計上の収入と支出に関する事項 （2） 品物の購入地域に関する事項 （3） 品物の購入先に関する事項 （4） 主要耐久消費財等に関する事項 （5） 年間収入及び貯蓄・借入金残高に関する事項 （6） 世帯及び世帯員に関する事項 （7） 現住居及び現住居以外の住宅・宅地に関する事項
調査時期	調査実施年の８月下旬～12 月上旬まで

＜使用したデータ概要＞

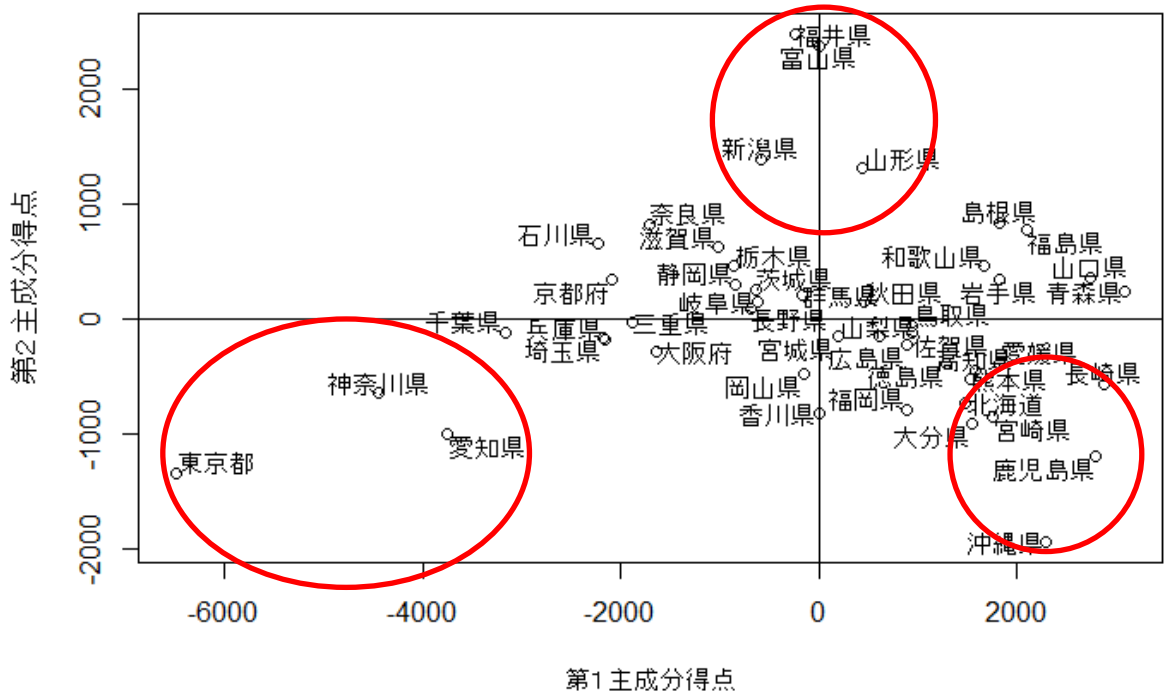
使用したのは、都道府県別の消費支出のうち食料に関する支出です。

左は、使用した統計表の一部です。右は、使用したデータ項目で、食品小分類の支出です。

収支項目	全国 Japan	都道府県 Prefectures 北海道 Hokkaido	青森県 Aomori-ken	岩手県 Iwate-ken	宮城県 Miyagi-ken	秋田県 Akita-ken
集計世帯数	54,208	2,164	739	746	789	770
世帯数分布（抽出率調整）	51,756,439	2,374,215	476,970	507,292	923,420	375,617
消費支出	254,402	221,757	215,712	220,630	251,694	229,200
食料	61,984	54,281	55,180	57,514	59,052	59,146
穀類	5,406	5,147	4,726	5,013	5,048	5,170
米	1,974	2,113	1,742	2,004	2,090	2,550
パン	2,049	1,760	1,617	1,633	1,632	1,296
麺類	1,098	1,017	1,163	1,111	1,114	1,173
他の穀類	285	258	204	264	213	152
魚介類	4,989	5,137	5,543	5,651	5,223	5,828
生鮮魚介	3,078	3,073	3,379	3,447	3,124	3,423
塩干魚介	824	923	1,056	1,022	902	1,189
魚肉練製品	487	434	371	474	563	385
他の魚介加工品	600	707	737	709	634	831
肉類	5,387	4,543	4,564	4,288	4,453	4,739
生鮮肉	4,346	3,468	3,513	3,358	3,454	3,721
加工肉	1,041	1,075	1,051	929	999	1,018
乳卵類	2,795	2,479	2,374	2,960	2,890	2,515
牛乳	1,041	915	922	1,112	1,112	998
乳製品	1,176	1,087	990	1,267	1,213	976
卵	579	477	461	581	564	541
野菜・海藻	7,108	6,401	6,572	7,241	7,379	7,046
生鮮野菜	4,959	4,441	4,465	4,795	4,959	4,786
乾物・海藻	499	426	496	533	520	530
大豆加工品	839	734	828	1,103	1,011	890
他の野菜・海藻加工品	810	800	783	811	889	840
果物	2,449	2,277	2,173	2,605	2,709	2,452
生鮮果物	2,319	2,157	2,054	2,514	2,575	2,362

- 穀類
  - 米
  - パン
  - 麺類
  - 他の穀類
- 魚介類
  - 生鮮魚介
  - 塩干魚介
  - 魚肉練製品
  - 他の魚介加工品
- 肉類
  - 生鮮肉
  - 加工肉
- 乳卵類
  - 牛乳
  - 乳製品
  - 卵
- 野菜・海藻
  - 生鮮野菜
  - 乾物・海藻
  - 大豆加工品
  - 他の野菜・海藻加工品
- 果物
  - 生鮮果物
  - 果物加工品
- 油脂・調味料
  - 油脂
  - 調味料
- 菓子類
- 調理食品
  - 主食的調理食品
  - 他の調理食品
- 飲料
  - 茶類
  - コーヒー・ココア
  - 他の飲料
- 酒類
- 外食
  - 一般外食
  - 学校給食

得られた結果は、次の図の通りです。



	第１主成分	第２主成分	第３主成分	第４主成分	第５主成分	第６主成分	第７主成分	第８主成分	第９主成分	第１０主成分
標準偏差 Standard deviation	2033.307	835.5115	709.2841	430.3122	352.0955	289.0146	263.0409	216.0203	164.4294	154.8071
寄与率 Proportion of Variance	0.6908	0.1167	0.08407	0.03094	0.02072	0.01396	0.01156	0.0078	0.00452	0.004
累積寄与率 Cumulative Proportion	0.6908	0.8075	0.89156	0.9225	0.94322	0.95718	0.96874	0.9765	0.98105	0.9851

この結果より、グラフ上、中心点周辺に位置する都道府県が多いことが分かりますが、一方で特徴的な分布となっている都道府県もあります。

① 第 1 主成分得点が低く（－）、第 2 主成分得点も低い（－）（グラフの左下に位置する）のは、東京、神奈川、愛知で、都市度が高く食料品の生産能力が高くない地域です。外食の支出が大きく、食料の支出も高めのグループです。

② 第 1 主成分得点が高く（＋）、第 2 主成分得点が低い（－）（グラフの右下に位置する）のは、沖縄、鹿児島、宮崎、大分、長崎で、九州地方の都道府県です。多様な食料品が生産されており、地産地消の消費傾向で食料の支出は低めのグループです。

③ 第 1 主成分得点は 0 付近、第 2 主成分得点が高い（＋）（グラフの中央上部に位置する）のは、福井、富山、新潟、山形です。北陸地方、東北地方の一部の地域であり、魚介や穀物類等の特定の食料品の生産が盛んなグループです。

第 1 主成分得点で寄与率は約 7 割であり、第 3 主成分得点までで累積寄与率は約 9 割になります。

Python、R での主成分分析

主成分分析のプロット図を作るために必要なコードは以下の通りです。

実施内容	各都道府県の食品小分類消費傾向に対する主成分分析
利用 Data	総務省　平成 26 年全国消費実態調査

Python コード（サンプル）

```
# 必要なライブラリのインポート
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
import os
from matplotlib.font_manager import FontProperties

# フォント等の設定
fp = FontProperties
  _(fname=r'C:¥WINDOWS¥Fonts¥YuGothR.ttc', size=14)

# データの読み込み
df = pd.read_csv("xxx.csv",encoding="SHIFT-JIS", index_col=0)
df = df.drop(columns=['食料'])

#主成分分析
pca = PCA()
pca.fit(df)
feature = pca.transform(df)

#　 主成分得点
# 利用する言語によって、 x 軸の方向が変わる場合がありますが結果に影響はありません。
plt.figure(figsize=(6, 6))
for (x, y, label) in zip(feature[:,0], feature[:,1], df.index):
    plt.scatter(x, y, alpha=0.8, c="b")
    plt.annotate(label, xy=(x,y), fontproperties=fp)
plt.grid()
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.show()
```

R コード（サンプル）

```
# データの読み込み
setwd("C:/xxx") #xxx 部分でフォルダを指定

dt1 <- read.csv("xxx.csv", header=T)
rownames(dt1)<-c("全国","北海道","青森県",
"岩手県","宮城県","秋田県","山形県","福島県",
"茨城県","栃木県","群馬県","埼玉県","千葉県",
"東京都","神奈川県","新潟県","富山県",
"石川県","福井県","山梨県","長野県","岐阜県",
"静岡県","愛知県","三重県","滋賀県","京都府",
"大阪府","兵庫県","奈良県","和歌山県",
"鳥取県","島根県","岡山県","広島県","山口県",
"徳島県","香川県","愛媛県","高知県","福岡県",
"佐賀県","長崎県","熊本県","大分県","宮崎県",
"鹿児島県","沖縄県")

# 主成分分析
(result<-prcomp(dt1[,c(3,4,5,6,7,8,9,10,11,12,13,
14,15,16,17,18,19,20,21,22,23,24,
25,26,27,28,29,30,31,32)],scale=F))
screeplot(result)
summary(result)
biplot(result)

# 主成分得点
# 利用する言語によって、 x 軸の方向が変わる場合がありますが結果に影響はありません。
result$x
plot(result$x[,1],result$x[,2],
xlab="第 1 主成分得点",
ylab="第 2 主成分得点")
abline(h=0)
abline(v=0)
library(maptools)
pointLabel(x=result$x[,1], y=result$x[,2], labels=dt1$都道府県)
```

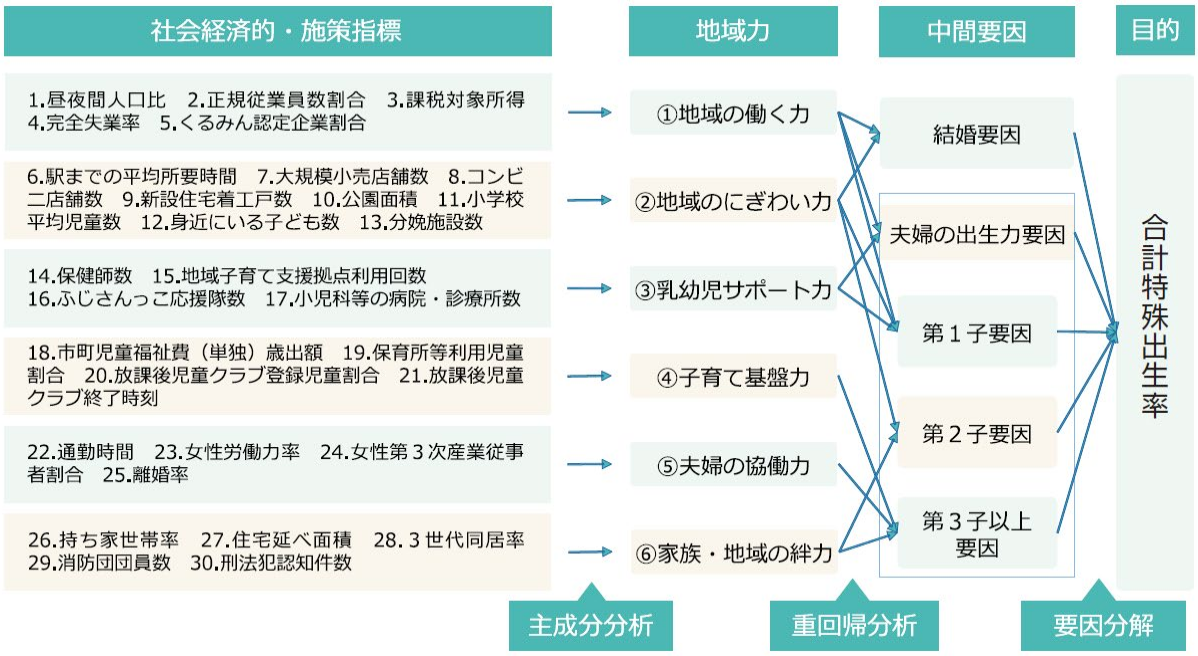
分析事例２：社会経済的･施策指標による地域力分析

次に、政策で実際に使われていた事例について紹介します。

【事例】ふじのくに少子化突破戦略事業（静岡県）

■ 静岡県では合計特殊出生率が、産業構造や立地条件により市町間に差が存在。県と市町が、結婚･妊娠･出産･育児の切れ目ない支援策を積極的に立案。

■ 合計特殊出生率を規定する5要因と、地域力を示す6要因との量的な関係を重回帰分析により把握。他方、地域力6要因は、主成分分析により30の社会経済的･施策指標をまとめている。これによりロジック･ツリーを定量化。



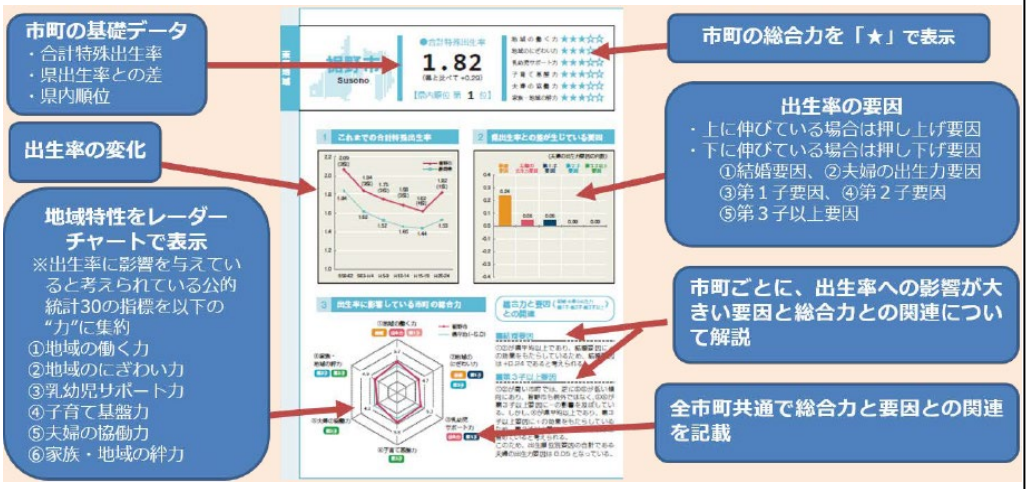
出典：ふじのくに（静岡県公式ホームページ）「ふじのくに少子化突戦略の羅針盤　合計特殊出生率「2」の達成に向けた基礎分析書　第２章　出生率に影響を与える要因の分析　p5」を引用

代表的手法として紹介した「主成分分析」と「重回帰分析」を課題抽出に用い、政策に応用した例です。

静岡県では、産業構造や立地条件の地域差が生じていること、また、同じ地域間でも市町別合計特殊出生率に差があることに注目し、35 市町の合計特殊出生率に関する要因分解を行い、地域別の課題を明らかにしました。統計的手法による課題抽出の試みです。

静岡県は市町及び有識者･市民･子育て支援団体等と協力し、30 個の社会経済的施策指標に関するデータを収集した上で主成分分析により縮約し、「地域力 6 要因」としました。これら 6 要因を、合計特殊出生率を規定する 5 要因と重回帰分析によってつなげ、合計特殊出生率を規定する 5 要因が「地域力 6 要因」のどの要因によって影響を受けているか、その影響の度合いはどの程度かを明らかにし、各市町別のカルテを作成しました。これにより、同じ県内とはいえ、市町別に対応するべき課題が異なることを定量的に明らかにし、各市町の少子化への独自取り組みを促進させました。

■ 静岡県の合計特殊出生率（の低さ）の要因がどこにあるかを市町別に可視化。



出所）静岡県「ふじのくに少子化突破戦略の羅針盤

出典：総務省統計局統計研修所 「政策立案と統計に係る応用的研修テキスト P57」より引用

このプロジェクトのポイントは、①仮説に基づいたデータを一から集めるのではなく、先行研究や国の計画等で既に使われている指標や e-Stat 等の政府統計情報を活用してデータを収集していること、また、②重回帰分析を行う際の仮説設定のため状況に応じた様々なレベルのエビデンス(証拠・根拠)を用いて定量的かつ統計的な分析に終始していることです。また、それぞれの主成分分析に使用するデータの設定の仕方などは参考になります。

また、のちの事業として、「ふじのくに少子化突破戦略応援事業費助成」を実施し、上記の分析結果を踏まえた、地域特性に応じた効果的な事業を実施する市町を支援する取り組みを行っています。統計的な手法の組み合わせにより、課題抽出を実施し、政策立案につなげた事例と言えます。

チャレンジ事項

ほかの公的統計データを用いて、主成分分析を実施しましょう。
その際に、主成分の累積寄与度と軸の数を増減させてみましょう。

（3）クラスター分析

本項における参考情報	✧　日本統計協会　統計学Ⅲ「多変量データ解析法　オフィシャル　スタディノート」

クラスター分析の概略

＜様々なクラスタリング法＞

クラスター分析のクラスターとは、英語で、「集団」、「群れ」を表す言葉です。クラスター分析とは、データの集まりから、互いに似た性質を持つデータを集め、それらをまとめてクラスターとして識別する、グループ分けのための分析方法です。

クラスター分析では、サンプル同士の「似ていること（類似度）」の基準、もしくは逆に「似ていないこと（非類似度）」の基準により、クラスタリングを行います。そのクラスタリングの方法には、1 つずつ類似度の高いサンプルを階層的にクラスタリングしていく階層的クラスタリング法と、はじめからいくつかの中心点を決め、その中心点に近い（非類似度が小さい）サンプルを一括してクラスタリングしていく非階層的クラスタリング法があります。

また、クラスターを作るのではなく、データ同士の類似性や違いを距離という概念を用いて表現する「多次元尺度構成法」という手法も存在しますが、ここでは多次元尺度構成法については割愛します。

クラスタリング法	目的	特徴
階層的クラスタリング	対象を分類すること	・直感的な解釈が可能。 ・事後的なクラスター数の解釈ができる。 ・対象数が大きくなると解釈が困難。
非階層的クラスタリング	対象を分類すること	・数値の読み取りが必要で直感的な解釈は難しい。 ・事前にクラスター数を決定する必要がある。 ・解釈に変数情報が利用できる。
多次元尺度構成法	対象間の関係を視覚化すること	・対象を低次元空間の点で表す。 ・類似性が大きいほど近くに付置する。 ・付置された次元を解釈する。

＜類似性データと非類似性データ＞

クラスタリング法でポイントとなるのは、データの種類がほかの手法と少し異なる部分です。具体的には（非）類似性データというデータを用います。

（非）類似性データとは、類似性データと非類似性データの両方を表すときに利用されます。類似性データは、対象間の似ている度合いを示し、非類似性データは対象間の似ていない度合いを示します。これらのデータは互いに変換が可能です。

類似データの例として、ソフトドリンクのブランドスイッチ、非類似データの例として都市間の距離の例を以下に記載します。例えば、類似性データとして、ソフトドリンクのブランドスイッチの例では、最初の時点であるソフトドリンクを購入した顧客が次の時点でどのソフトドリンクを購入しているかを表しています。最初の時点（T 時点）でコーラを買っていて、次の時点（T+1 時点）で紅茶を選んだ人は、70 人という読み方ができます。

非類似性データとして、都市間の距離データの例では、A 市から B 市の距離は 1500（km）という読み方ができます。

類似性データ		T＋1時点			
		コーラ	紅茶	･･･	水
T時点	コーラ	100	70	･･･	10
	紅茶	60	150	･･･	5
	･･･	･･･	･･･	･･･	･･･
	水	10	40	･･･	180

非類似性データ	A市	B市	･･･	N市
A市	0	1500	･･･	2000
B市	1500	0	･･･	3000
･･･	･･･	･･･	･･･	･･･
N市	2000	3000	･･･	0

また、（非）類似性データは多変量データを変換しても得ることができます。各変量間の相関係数を計算し、それを各成分として持つ相関行列を作れば、類似性データへと変換することが可能です。

多変量データ		データ内容			
		体育館の数	高校の数	公園の数	･･･
地域ID	001	10	7	15	･･･
	002	6	2	8	･･･
	･･･	･･･	･･･	･･･	･･･
	047	13	4	10	･･･

変換

相関行列化				
類似性データ	体育館の数	高校の数	公園の数	･･･
体育館の数	1.00	0.71	0.50	･･･
高校の数	0.71	1.00	0.32	･･･
公園の数	0.50	0.32	1.00	･･･
･･･	･･･	･･･	･･･	1.00

また、多変量データを非類似性データに変換する場合には、各レコードを N 次元空間（N はデータ内容数）上の点として捉え、それぞれの点の間のユークリッド距離を計算すると計算された距離を各要素に持つようなデータは非類似性データとなります。下図では、世帯人員構成が似ていれば似ているほど距離は小さくなり、違っていれば違うほど距離は大きくなります。

多変量データ		世帯人員別一般世帯数(総務省,2010)（単位：千世帯）			
		1人	2人	3人	・・・
地域ID	北海道	843	768	418	・・・
	青森	141	143	99	・・・
	岩手	132	129	91	・・・
	・・・	・・・	・・・	・・・	・・・

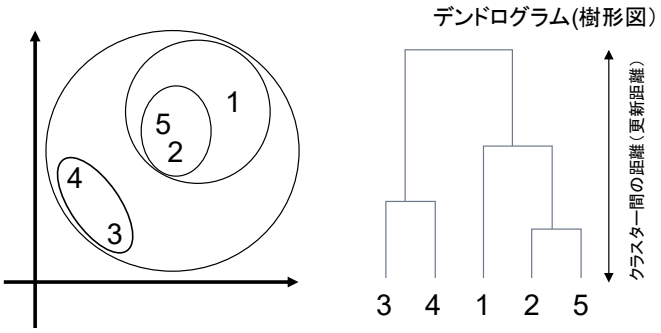
変換

ユークリッド距離化				
非類似性データ	北海道	青森	岩手	・・・
北海道	0	1015.51	1033.40	・・・
青森	1015.51	0	19.82	・・・
岩手	1033.40	19.82	0	・・・
・・・	・・・	・・・	・・・	0

第2講

＜階層的クラスタリング法＞

階層的クラスタリング法の具体的な手続きについて説明します。階層的クラスタリング法では、逐次的に対象を結合し、クラスターを階層的に構成することを目的とします。



上記の図をデンドログラム（樹形図）と呼びます。階層的クラスタリング法は、最も似ている性質のデータの組み合わせから順にクラスターにしていく方法ですが、クラスターにしていく順を階層で表した図です。デンドログラムから解るように、一番近い 2 と 5 が最初に結合されてクラスターになり、その後に 3 と 4 が結合して次のクラスターができ、2 と 5 と 1 が結合してクラスターができ、最後にそれらが全て結合するという値を示しています。もしもすべての点が 2 次元上に表されたとすると、上記、左側の図のようになります。

階層的クラスタリングの手順は次の 4Step となります。

Step		内容
1	クラスター間距離の算出	全てのクラスターの組に対して、クラスター間距離を求める。
2	クラスターの結合	クラスター間距離が最小なクラスターの組を結合し、新たなクラスターを作成する。
3	クラスター間距離の更新	新たなクラスターとその他のクラスターとのクラスター間距離を求める。
4	繰り返し	クラスター数があらかじめ決められた値（通常は 1）になるまで Step2,3 を繰り返す。

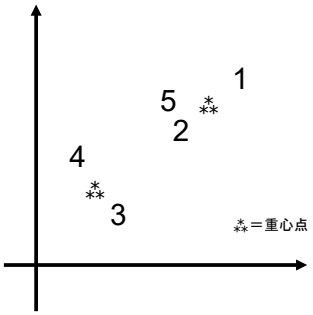
階層的クラスタリング法は、クラスター間距離の定義によって手法が異なってきます。具体的な内容には触れませんが、階層的クラスタリング法としては、最短距離法、最小距離法、群平均法、重心法、ウォード法などがあります。

最後に、階層的クラスタリング法のメリット・デメリットは以下の通りです。

メリット	デメリット
・結果が視覚的に表現され、解釈が容易である。 ・クラスター数を事前に定める必要が無く、クラスタリング過程を解釈することができる。 ・多変量データ、（非）類似性データの両方に適用可能である。 ・量的データのほか、質的データでも使える。	・対象が多い場合は解釈が困難となる。 ・クラスターの解釈に対象の情報を必要とする。 ・手法により、クラスタリング結果が大きく異なる。

＜非階層的クラスタリング法＞

非階層的クラスタリング法の中でも最も一般的な k-means 法について説明します。k-means 法の目的は、あらかじめ定めたクラスター数で各クラスター内のばらつき(k)を最小にするようなクラスターを求めることです。例えば、データが 2 変量のデータで与えられたときは、以下のような図になります。



k-means 法で結果として得られるのは、各対象の所属しているクラスターとそれぞれのクラスターの重心です。

非階層的クラスタリングの手順は次の 5Step となります。

Step		内容
1	クラスター数、クラスター中心の初期設定	クラスター数を定め、クラスター中心の初期値を定める。
2	クラスターの更新	最短距離基準に基づき、全ての対象（全データ）を最も近いクラスターに分属させる。もし、対象から非類似性が最も近いクラスター中心が複数存在する場合には、適当（同じ距離のクラスターから無作為）に 1 つ選択して、分属させる。
3	クラスター中心の更新	全てのクラスター中心を再計算する。
4	目的関数の値を計算	目的関数 L の値を計算する。 k-means 法の目的関数は、各クラスターのばらつきの和を示しています。
5	繰り返し	目的関数の値が変わらなくなるまで、Step2,3,4 を繰り返す。

最後に、非階層的クラスタリング法のメリット・デメリットは以下の通りです。

メリット	デメリット
・対象の数が多い場合も適用可能である。 ・解析結果が安定している。 ・クラスターの解釈に変量の情報が利用可能である。	・事前にクラスター数を定める必要がある。 ・解釈には数値を読み取る必要があり、直感的には解釈できない。

クラスタリング法には様々な種類があり、データの変換方法も多様に存在します。方法が異なればクラスタリング結果も異なるため、クラスタリング法を使う際には最低限以下のことに注意してください。

・各方法の特徴をきちんと理解すること。
・いくつかの方法で分析して結果を比較すること。
・利用した手法を明記すること。
・元のデータを変換したデータを用いる場合はその変換についても明記すること。

分析事例：都道府県別食料品消費傾向

上記の非階層的クラスタリング法の k-means 法を用いた分析例を紹介します。

利用するデータは総務省が実施している「全国消費実態調査」という公的統計データです。このデータを用いて、都道府県を分けるクラスター分析を行い、その結果を解釈しましょう。

＜使用した統計の特徴＞

全国消費実態調査（2019 年からは全国家計構造調査）を使用しました。詳細は、主成分分析の事例 1 でも同じ統計を使用しているため、そちらを参照ください。

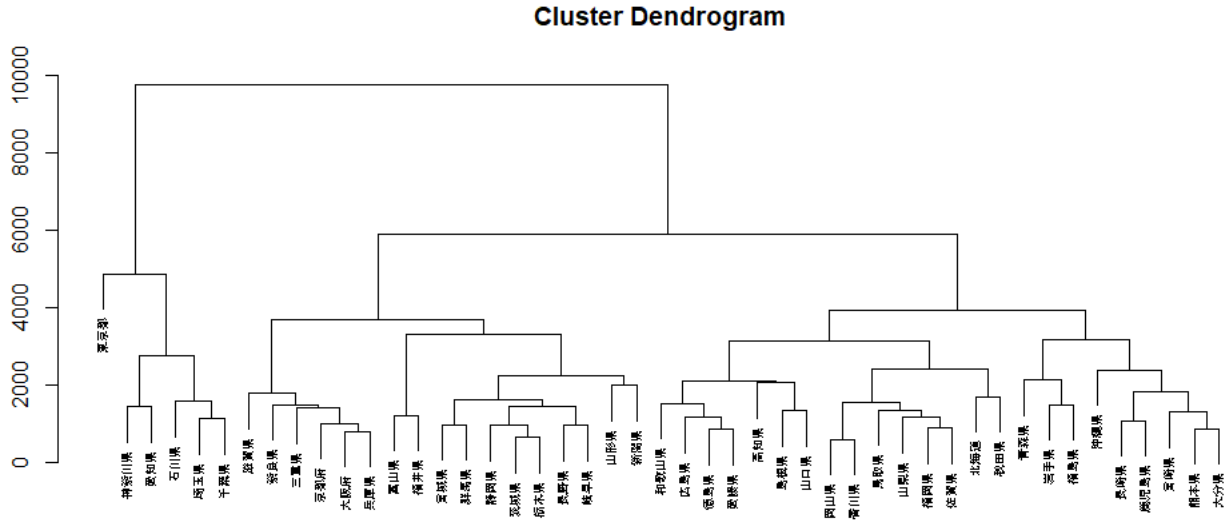
＜使用したデータ概要＞

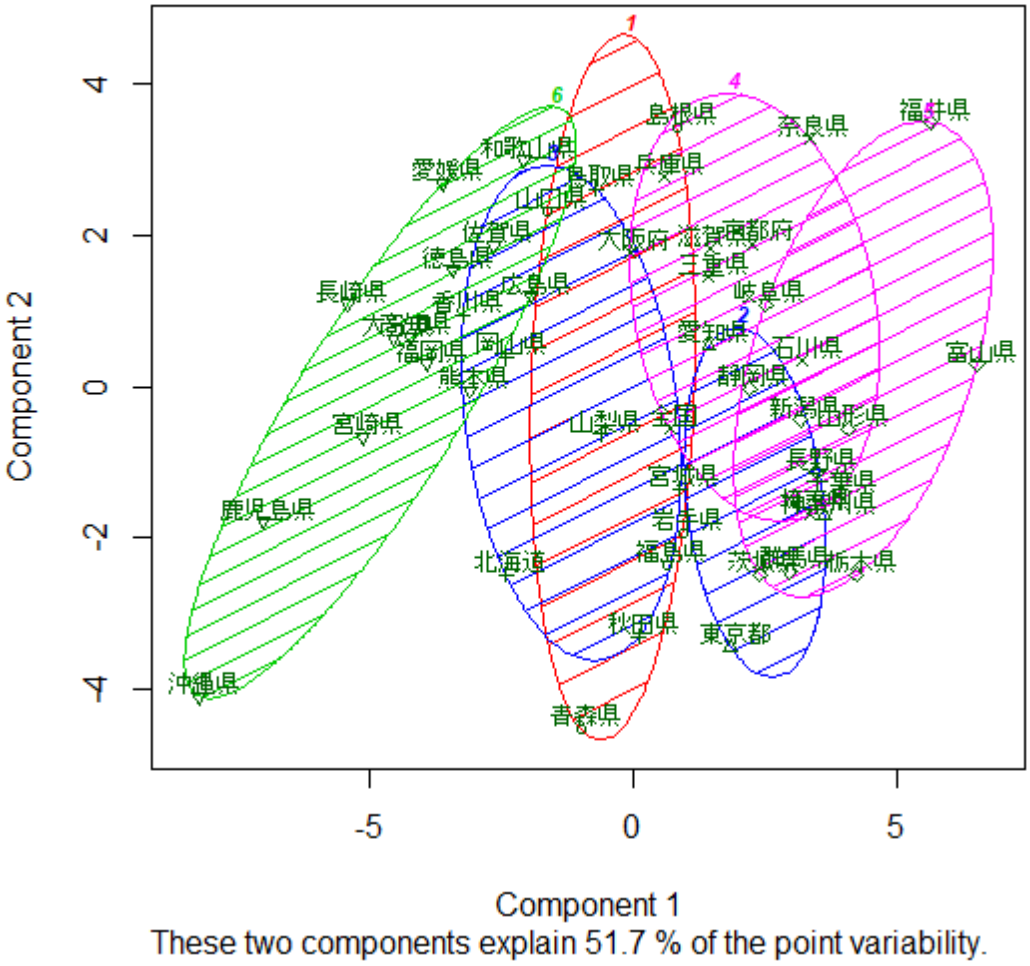
消費食料小分類項目別データ（多変量データ）を使用しました。

＜分析の目的＞

各都道府県世帯別の特徴を捉え、上述の通り、多変量データ間の相関行列を作成し、類似性データを作り、都道府県別の類似度から階層的なクラスタリングを行います。

得られたデンドログラムと、グループ図が下図となっています。





上記の階層型クラスタリングの方法によるデンドログラムの結果から最も特徴的なのは、東京都が他の道府県と比べて異なる消費性向があることがわかります。また、「埼玉県と千葉県」、「大阪府、兵庫県と京都府」、「長野県と岐阜県」など隣接する都道府県で同じ傾向にあることや、九州の「長崎県、鹿児島県、宮崎県、熊本県、大分県」は近い傾向にあることなどの特徴が見ることができます。

また、非階層型のクラスタリングの方法によると、第 2 グループで関東と中部地方の都市型の傾向を持つ都道府県が同一のクラスターを作成し、第 5 グループでは、東日本の農産物の生産力の高い地域、第 6 グループでは西日本の農産物の生産力の高い地域が同一のクラスターを作成していることがわかります。

Python、R でのクラスター分析

クラスター分析を行うために必要なコードは以下の通りです。

実施内容	各都道府県の食品小分類消費傾向に対するクラスター分析
利用 Data	総務省　平成 26 年全国消費実態調査

Python コード（サンプル）

```
#必要なライブラリの読み込み
import pandas as pd
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import linkage, dendrogram
import codecs
from matplotlib.font_manager import FontProperties
fp = FontProperties(fname=r'C:xxx.ttc', size=14)
from sklearn.datasets import make_blobs
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA

#  xxx.csv 部分で作業 csv を指定する
with codecs.open("xxx.csv", "r", "Shift-JIS", "ignore") as file:
    df = pd.read_table(file, delimiter=",")
df = df.set_index("都道府県")
df.iloc[:, 1:32].head()

# クラスター分析 1　デンドログラムの描画
result1 = linkage(df.iloc[:, 1:32],
                  metric = "euclidean",
                  method = "complete")
dendrogram(result1,labels=df.index )
for l in plt.gca().get_xticklabels():
    l.set_fontproperties(fp)
    plt.title("Dedrogram")
plt.ylabel("Threshold")

# クラスター分析 2
km = KMeans(n_clusters=6, n_jobs=2)
model=km.fit(df[df.columns[1:32]])
feature = km.transform(df[df.columns[1:32]])

# クラスタの散布図
color = ["red","pink","green","orange","blue","yellow"]
plt.figure(figsize=(6, 6))
for (x, y, label) in zip(feature[:,0], feature[:,1], df.index):
```

```
    plt.scatter(x, y, alpha=0.8, c="b")
    plt.annotate(label, xy=(x,y), fontproperties=fp)

for i in range(feature.shape[0]):
    plt.scatter(feature[i, 0], feature[i, 1],c=color[int(model.labels_[i])])
plt.show()
```


R コード（サンプル）

```
install.packages("maptools")
# C:xxx 部分で作業フォルダを指定する
setwd("C:xxx ")

# xxx.csv 部分で作業 csv を指定する
dt1 <- read.csv("xxx.csv", header=T)
rownames(dt1)<-c("北海道","青森県","岩手県","宮城県","秋田県","山形県","福島県","茨城県","栃木県","群馬県","埼玉県","千葉県","東京都","神奈川県","新潟県","富山県","石川県","福井県","山梨県","長野県","岐阜県","静岡県","愛知県","三重県","滋賀県","京都府","大阪府","兵庫県","奈良県","和歌山県","鳥取県","島根県","岡山県","広島県","山口県","徳島県","香川県","愛媛県","高知県","福岡県","佐賀県","長崎県","熊本県","大分県","宮崎県","鹿児島県","沖縄県")

# クラスター分析 1　デンドログラムの描画
res <- hclust(dist(dt1[,c(3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32)]))
plot(res,labels=dt1$都道府県,cex=0.5)

# クラスター分析 2
res2 <- kmeans(dt1[,c(3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32)],centers=6)
library(cluster)
clusplot(dt1[,c(3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32)], res2$cluster,
color=TRUE, shade=TRUE, ,main="",lines=0 ,labels=2,cex=0.7)

# クラスタの散布図
png(filename="クラスター分析 2.png", width=2339, height=1654, pointsize = 40) ####  200 dpi
clusplot(dt1[,c(3,4,5,6,7,8,9,10,11,12,13,14)], res2$cluster, color=TRUE, shade=TRUE,main="", lines=0 ,labels=2,cex=0.7)
dev.off()
```