Stata19 whitepapers

mwp-111

## cluster linkage - 階層型クラスタ分析 【 評価版 】

Stata の cluster (及び clustermat) コマンドは分割型と並んで階層型のクラスタ分析機能をサポートして いますが、それは本 whitepaper に記す linkage コマンド群によって提供されます。

1. 階層型クラスタ分析
 2. 階層クラスタリングの用例

Example 1 Example 2 Example 3

## 1. 階層型クラスタ分析

Stata がサポートする階層型クラスタ分析の機能は凝集型 (agglomerative) と呼ばれるタイプに属するわけで すが、グループ間の比較に用いられる手法 — 連結法 (linkage method) — に種々のものがあるため、それぞ れの手法ごとに別個のコマンドが対応する形となっています。具体的なコマンドとしては次の7種類がありま す。なお、これらの連結法の特質については [MV] cluster (*mwp-110*) をご参照ください。

コマンド	機能
cluster singlelinkage	単連結 (single linkage) 法によるクラスタ分析
cluster averagelinkage	平均連結 (average linkage) 法によるクラスタ分析
cluster completelinkage	完全連結 (complete linkage) 法によるクラスタ分析
cluster waveragelinkage	加重平均連結 (weighted-average linkage) 法によるクラスタ分析
cluster medianlinkage	メディアン連結 (median linkage) 法によるクラスタ分析
cluster centroidlinkage	重心連結 (centroid linkage) 法によるクラスタ分析
cluster wardslinkage	Ward 連結 (Ward's linkage) 法によるクラスタ分析

<sup>ⓒ</sup> Copyright Math 工房; 一部 ⓒ Copyright StataCorp LP (used with permission)

## 2. 階層クラスタリングの用例

▷ Example 1: 単連結法

[MV] cluster linkage の Example 1 には単連結法を用いた階層クラスタリングの例が記載されています。 使用するのは Example データセット labtech.dta です。

. use https://www.stata-press.com/data/r19/labtech.dta  $^{\ast1}$ 

このデータセット中には熱帯雨林で収集された同一種と見られる植物標本 50 体に対する計測データが記録されています。変数である x1-x4 は計測に用いられた化学物質の種類を表しています。次に示すのは先頭 10 件のデータです。

. list in 1/10  $^{\ast 2}$ 

	x1	x2	x3	x4	labtech
1.	17.4	78.6	101.4	109.3	Jen
2.	87	30.1	79.1	6.6	Jen
3.	.1	.6	.9	.2	Sam
4.	106	10	44.6	57.6	Deb
5.	140.6	122	114.8	122.8	Jen
					·····
6.	.5	.8	.2	.5	Sam
7.	140.5	27	71.2	69.6	Bill
8.	.5	.3	.5	.7	Sam
9.	18	74.4	26.4	116	Al
10.	3.6	143.6	12.8	70.1	Deb

これらの標本はすべて同一種の植物であると想定されていますが、その中にサブグループが存在しないか、あるいは他と異なる特異な標本が存在しないかを見るためにクラスタ分析を実行してみます。連結法としては単連結法を、距離の尺度としてはデフォルトである L2(ユークリッド距離)<sup>\*3</sup>を用いることにし、分析結果には sngeuc という名称を付けることにします。

- Statistics ▷ Multivariate analysis ▷ Cluster analysis ▷ Cluster data ▷ Single linkage と操作
- Main タブ: Variables: x1 x2 x3 x4

(Dis)similarity measure: Continuous (デフォルト)

L2 or Euclidean (デフォルト)

Name this cluster analysis: sngeuc

<sup>\*1</sup> メニュー操作: File ▷ Example Datasets ▷ Stata 19 manual datasets と操作、Multivariate Statistics Reference Manual [MV] の cluster linkage の項よりダウンロードする。

<sup>\*&</sup>lt;sup>2</sup> メニュー操作: Data ▷ Describe data ▷ List data

<sup>\*&</sup>lt;sup>3</sup> 類似度/非類似度尺度については [MV] cluster (*mwp-110*) を参照。

📃 clust	ter singlelii	nkage - Single-linkage	e clustering	—		×
Main	if/in	Advanced				
Variab	les: (leave	empty for all)				
x1 x2	x3 x4					$\sim$
(Dis)s	similarity n	leasure				
00	ontinuous	OBinary	Mixed			
L2 o	or Euclidea	n .				
L2so	quared or s	quared Euclidean	1			
L1 o	or absolute	value vimum value				
	inty of ma					
2	# 604					
2	# 10r	L(#) of Lpower(#)				
Name	this cluster	analysis: (leave emp	ty for default name)			
sngeu	с					
? C			ОК	Cancel	Su	bmit
	<b>2</b>	l cluster singl	elinkage ダイアログ -	Main タ	ブ	

. cluster singlelinkage x1 x2 x3 x4, measure(L2) name(sngeuc)

## 評価版では割愛しています。

cluster dendrogram コマンドを使用するとクラスタリングに関するデンドログラム(樹形図)を作成する ことができます。

- Statistics ▷ Multivariate analysis ▷ Cluster analysis ▷ Postclustering ▷ Dendrograms と操作
- Main タブ: Cluster analysis: sngeuc

cluster dendrogram - Dendrograms for hierarchical cluster analysis								—	×		
Main	if/in	Plot	Add plots	Y axis	X axis	Titles	Legend	Overall			
Cluster sngeu	Cluster analysis: Sngeuc Update Plot the full dendrogram Plot top branches only Plot branches above a (dis)similarity measure										
Do not center parent branches											
Variable containing leaf labels:											
Orier	ntation: ertical orizontal										

図 3 cluster dendrogram ダイアログ - Main タブ

・X axis タブ: Major tick/label properties: Labels タブ: Size: Small

Angle: 90°

```
. cluster dendrogram sngeuc, xlabel(, labsize(small) angle(ninety))
invalid line in style file anglestyle-ninety:
*label "90 degrees"
```



このデンドログラムから見て取れることはまず最初に、横軸中央部に位置する観測データ群が短い縦棒の長さ で示されているように互いに近接しており、長い縦棒を有する他の観測データからは大きく隔たっているとい うことです。次にこれらの10個の観測データを除外して考えるなら、残されたデンドログラムからは明確な クラスタリングが読み取れないことがわかります。これはデンドログラム上部における縦棒の長さがみな比較 的短いことに示されています。

評価版では割愛しています。

▷ Example 2: 2 値データ

評価版では割愛しています。