クラスター分析（階層型）

はじめに

cluster コマンドまたは **clustermat** コマンドはリンク手法を指定して、階層的集約クラスター分析を実行します。リンク手法に関しては、さまざまな手法があり詳細については **cluster** コマンドのマニュアルをご参照ください。

cluster コマンドまたは **clustermat** コマンドを実行後に、**cluster dendrogram** コマンドを実行してデンドログラムを表示することができます。また、**cluster stop** コマンドまたは **clustermat stop** コマンドはグループ数を決定する際に役立ちます。さらに、**cluster generate** コマンドを使用してグループ変数を生成できます。

例題 1:

小さなバイオテクノロジー企業に勤めるシニアデータアナリストに熱帯雨林から収集された特定の植物の 50 の異なるサンプルに関する 4 つの化学実験室測定値を含むデータセットが提供されました。サンプルを収集した責任者は、原住民からの情報に基づいて、植物からの抽出物が、会社でベストセラーの栄養補助食品に関連する負の副作用を軽減する可能性があると考えています。

会社の化学者と植物学者が植物の可能な使用法を検討し、実験を計画している間、予備データをもとに、研究の役に立つような情報を報告するよう依頼されました。

50 の植物はすべてが同じタイプであると想定されていますが、クラスター分析を実行して、それらの間にサブグループまたは異常があるかどうかを確認することにします。初期設定のユークリッド距離でシングルリンケージのクラスター分析を行うことにしました。

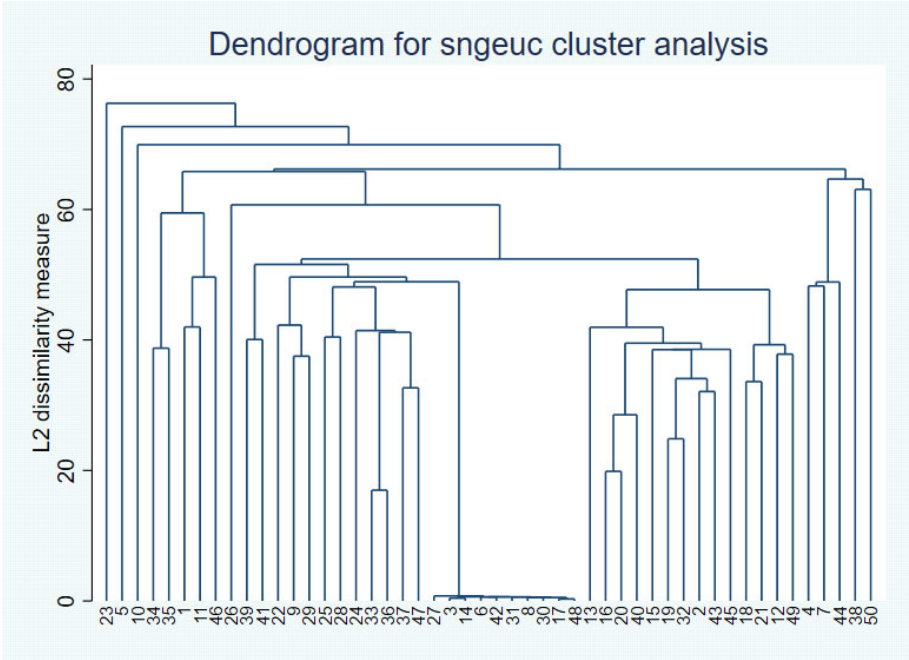
```
. use https://www.stata-press.com/data/r17/labtech
. cluster singlelinkage x1 x2 x3 x4, name(sngeuc)
. cluster list sngeuc
```

```
sngeuc  (type: hierarchical,  method: single,  dissimilarity: L2)
      vars: sngeuc_id (id variable)
            sngeuc_ord (order variable)
            sngeuc_hgt (height variable)
     other: cmd: cluster singlelinkage x1 x2 x3 x4, name(sngeuc)
            varlist: x1 x2 x3 x4
            range: 0 .
```

cluster singlelinkage コマンドは、いくつかの変数と sngeuc という名前のクラスターオブジェクトを生成します。cluster list コマンドはクラスターオブジェクトに関する情報を提供します。

次に、デンドログラムを確認します。

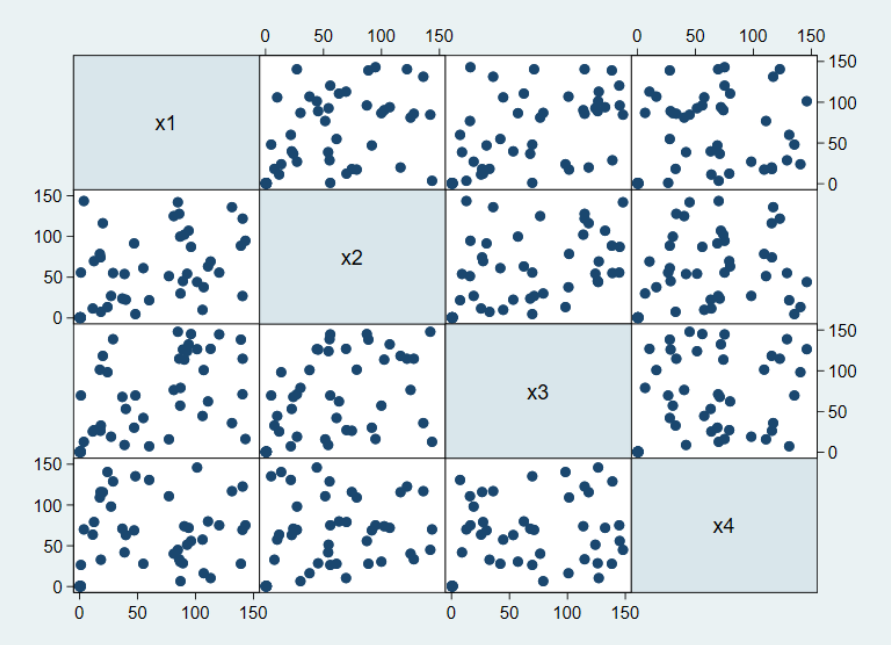
```
. cluster dendrogram sngeuc, xlabel(, angle(90) labsize(*.75))
```



これまでの分析経験を踏まえて、2 つの考察が得られました。1 つ目はデンドログラムの中央部分のいくつかの観測値は近く（垂線が短い）、他の観測値とは離れている（垂線が長い）ことがわかります。次に、中央部分の 10 の観測値を除くと、デンドログラムの上部にある比較的短い垂線で示されているように明確なクラスターは存在しません。

10 の観測値が近い理由を探るため、散布図を作成します。次のコマンドで散布図行列を表示します。

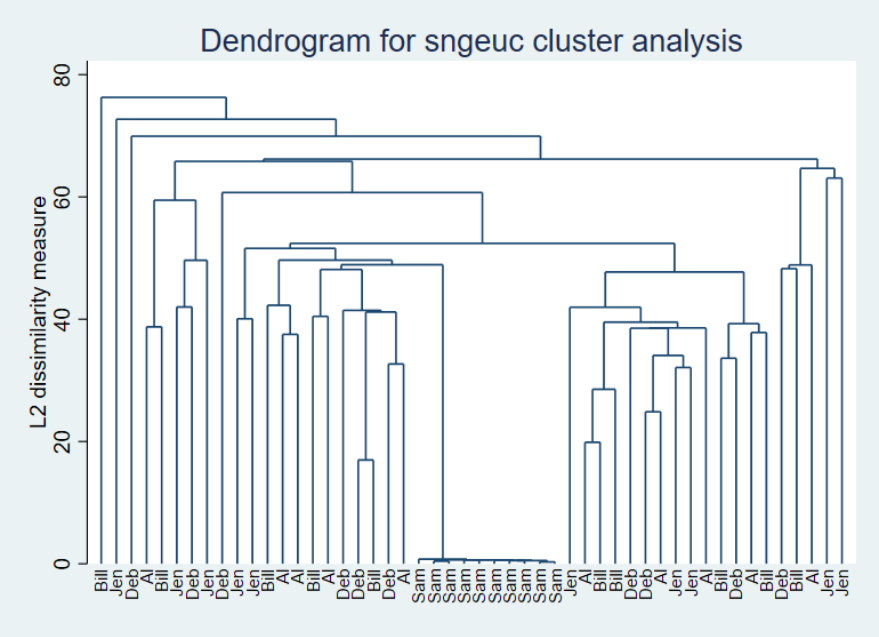
```
. graph matrix x1 x2 x3 x4
```



表示された散布図からは特段の考察は得られません。

しかし、ここであることを思い出します。過去にデータの取り扱いに関する事故があり、会社は各データセット内に測定を行った技術者の名前を与える変数を配置するという防止策を規則として整備していました。そこで、技術者の名前をラベルとしてデンドログラムを表示することにしました。

```
. cluster dendrogram sngeuc, labels(labtech) xlabel(, angle(90) labsize(*.75))
```



予期した通り、中央部分の 10 の観測値はサムという分析者が測定を行ったデータであることがわかりました。データの一覧を確認し、他の 4 人の技術者のデータ範囲は最大約 150 ですが、サムのデータは 0 以上 1 以下となっていました。サムは、サンプルを分析する前にセンサーを調整することを忘れたようです。このクラスター分析の結果をまとめ、データを修正するよう研究室に送りました。

例題 2:

学部生の講義を担当する社会学の教授から、課題として 30 の被験者に関する 30 行 60 列の二値変数のデータを考察する宿題が出されました。この課題は取り組みづらいなと思いつつも、単位を取得するため分析を始めました。

数ある分析の中で、以下のクラスター分析に挑戦します。理解のしやすさを考慮して、単純一致係数を用いたシングルリンケージクラスタリングを使用することにします。また、generate オプションを指定して、zstub という接頭辞をもつ変数を生成します。name オプションを指定しない場合、Stata がクラスター分析に名前を付与します。

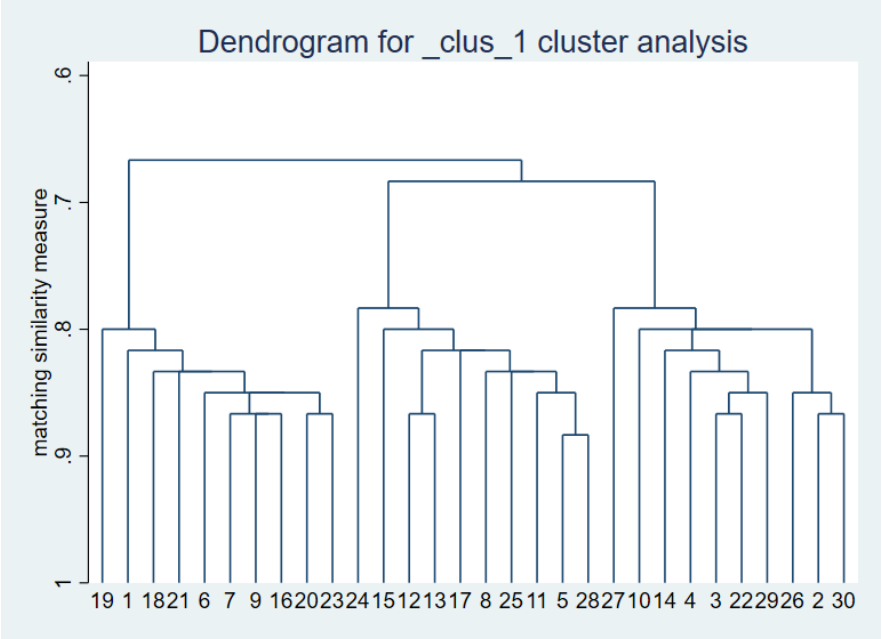
```
. use https://www.stata-press.com/data/r17/homework, clear
. cluster s a1-a60, measure(matching) gen(zstub)
. cluster list
```

```
_clus_1  (type: hierarchical,  method: single,  similarity: matching)
      vars: zstub_id (id variable)
            zstub_ord (order variable)
            zstub_hgt (height variable)
     other: cmd: cluster singlelinkage a1-a60, measure(matching) gen(zstub)
            varlist: a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 a11 a12 a13 a14 a15 a16 a17
                 a18 a19 a20 a21 a22 a23 a24 a25 a26 a27 a28 a29 a30 a31 a32
                 a33 a34 a35 a36 a37 a38 a39 a40 a41 a42 a43 a44 a45 a46 a47
                 a48 a49 a50 a51 a52 a53 a54 a55 a56 a57 a58 a59 a60
            range: 1 0
```

Stata はクラスター名を_clus_1 と設定し、変数 zstub_id, zstub_ord, zstub_hgt を生成しました。

cluster tree コマンドを実行してデンドログラムを表示します。このデンドログラムは cluster dendrogram コマンドを実行しても表示されます。Stata は初期設定で直前に実行されたクラスター分析を参照するため、名前を入力する必要はありません。

```
. cluster tree
```



デンドログラムより、30 行の観測行の中に 3 つのグループが存在するように思われます。おそらく教授はこの構造を見つけてほしいと思っていると考え、レポートを書き始めました。3 つのグループについてさらに調べるために、cluster generate コマンドを実行してグループ変数を生成しました。様々な記述統計を確認し、3 グループの集計を行ってレポートを書き上げました。

課題を提出した後、教授から truegrp という変数が追加された同じデータセットを与えられました。変数 truegrp は教授の考えたグループを示しているようです。そこで、課題に対する評価を確認するため、グループ変数 grp3 を作成して、truegrp と grp3 のクロス集計を行いました。

```
. cluster generate grp3 = group(3)
. table grp3 truegrp, nototals
```

	truegrp		
	1	2	3
grp3			
1		10	
2			10
3	10		

3 つのグループへの任意の数字の割当て差はありますが、提出した課題と教授の回答は一致しているようです。課題を達成することができたと知り、安心しました。

シングルリンケージクラスタリングに加えて、メディアンリンケージの結果を確認してみましょう。シンプルリンケージクラスタリングと同様にグループ間の類似性尺度には単純一致係数を使用します。name オプションを指定して、このクラスター分析に medlink という名前を付与します。cluster list コマンドで詳細を表示します。

```
. cluster median a1-a60, measure(match) name(medlink)
. cluster list medlink
```

```
medlink  (type: hierarchical,  method: median,  similarity: matching)
      vars: medlink_id (id variable)
            medlink_ord (order variable)
            medlink_hgt (real_height variable)
            medlink_pht (pseudo_height variable)
     other: cmd: cluster medianlinkage a1-a60, measure(match) name(medlink)
            varlist: a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 a11 a12 a13 a14 a15 a16 a17
                 a18 a19 a20 a21 a22 a23 a24 a25 a26 a27 a28 a29 a30 a31 a32
                 a33 a34 a35 a36 a37 a38 a39 a40 a41 a42 a43 a44 a45 a46 a47
                 a48 a49 a50 a51 a52 a53 a54 a55 a56 a57 a58 a59 a60
            range: 1 0
```

cluster dendrogram コマンドを実行してデンドログラムを表示しようとしましたが、このクラスター分析では反転が生じているためデンドログラムは描けませんでした。反転が生じると結果を容易に解釈することができなくなります。

cluster generate コマンドを実行して 3 グループのグループ変数を生成し、truegrp

と比較します。

```
. cluster generate medgrp3 = group(3)
. table medgrp3 truegrp, nototals
```

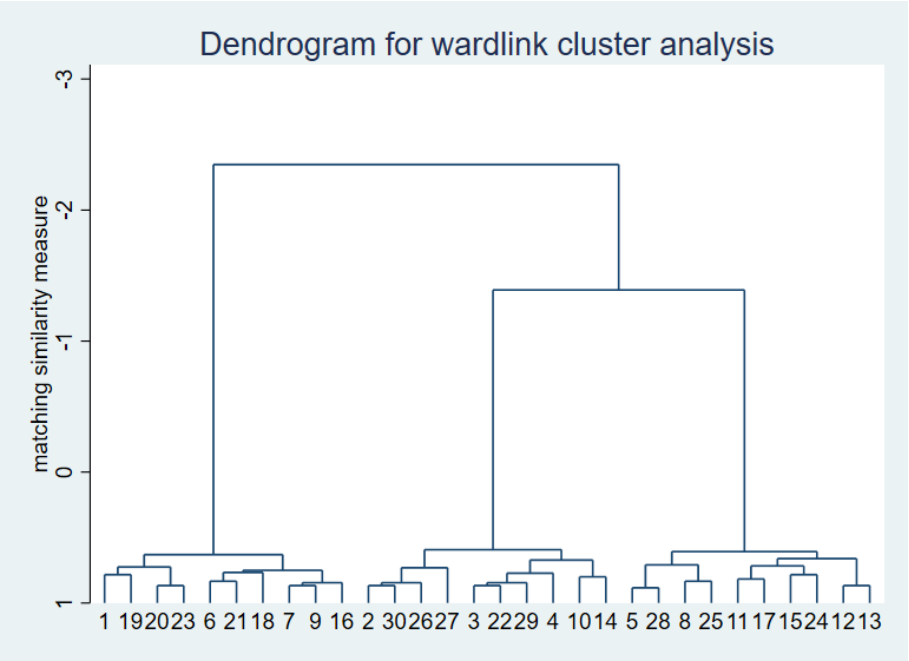
	truegrp		
	1	2	3
medgrp3			
1		10	
2	10		
3			10

メディアンリンケージクラスタリングを使用したことでデンドログラムを表示できないため、クラスタリング手法をウォードリンケージに変えて実行します。

```
. cluster ward a1-a60, measure(match) name(wardlink)
. cluster list wardlink
```

```
wardlink  (type: hierarchical,  method: wards,  similarity: matching)
      vars: wardlink_id (id variable)
            wardlink_ord (order variable)
            wardlink_hgt (height variable)
     other: cmd: cluster wardslinkage a1-a60, measure(match) name(wardlink)
            varlist: a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 a11 a12 a13 a14 a15 a16 a17
                 a18 a19 a20 a21 a22 a23 a24 a25 a26 a27 a28 a29 a30 a31 a32
                 a33 a34 a35 a36 a37 a38 a39 a40 a41 a42 a43 a44 a45 a46 a47
                 a48 a49 a50 a51 a52 a53 a54 a55 a56 a57 a58 a59 a60
            range: 1 0
```

```
. cluster tree wardlink
```



シングルリンケージクラスタリングのように、ウォードリンケージクラスタリングのデンドログラムからも 30 の観測値の中に 3 つのグループが存在しているように思われます。しかし、y 軸の範囲を確認すると類似度係数が 1 から-2 と-3 の間となっています。定義によると、一致係数は 0 以上 1 以下となります。これはウォードリンケージクラスタリングを用いたために生じ、尺度の選択に関する議論において警告されています。詳細は、[MV]cluster 内の Dissimilarity transformations and the Lance and Williams formula と Warning concerning similarity or dissimilarity choice を参照してください。

truegrp と wardgrp3 のクロス集計を表示します。

```
. cluster generate wardgrp3 = group(3)
. table wardgrp3 truegrp, nototals
```

	truegrp		
	1	2	3
wardgrp3			
1		10	
2	10		
3			10

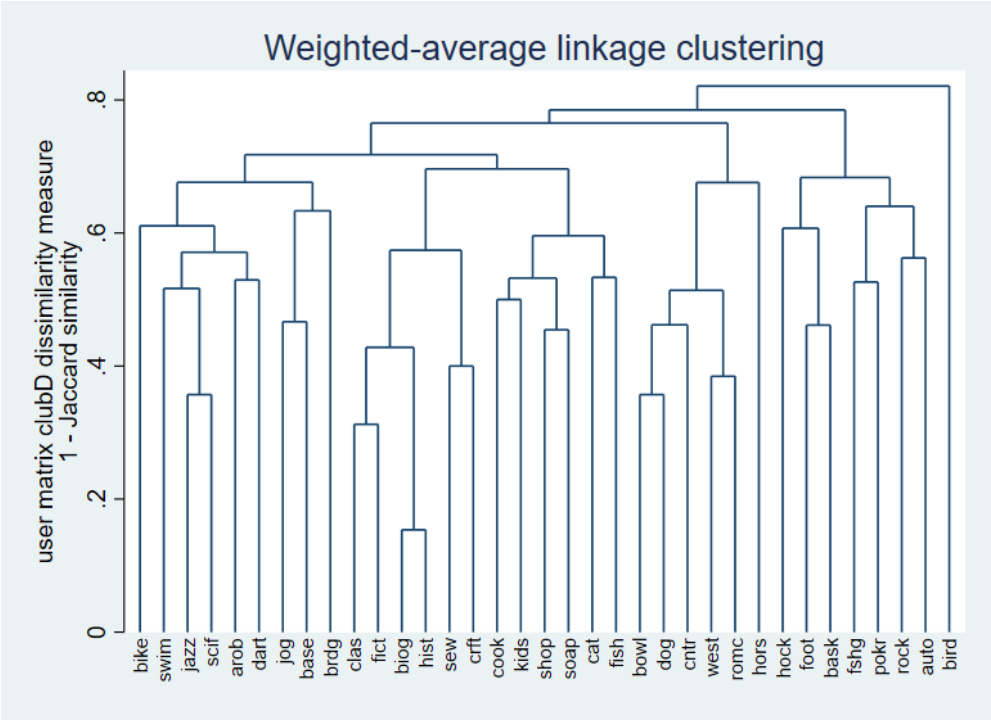
3 つのグループへの任意の数字の割当て差はありますが、ウォード法の結果と教授の回答は一致しているようです。ウォードリンケージとユークリッド距離の二乗以外の尺度を使用しましたが、解釈可能なクラスター分析の結果を得ることができました。

例題 3:

wclub データセットには 30 名の女性から得た 35 問の二値（yes or no）回答が含まれています。加重平均リンケージクラスタリングによって（観測行ではなく）35 の変数がどのようなクラスタを生成するか分析を行います。

Matrix dissimilarity コマンドを使用して、1-Jaccard 係数に等しい非類似度行列を生成します。

```
. use https://www.stata-press.com/data/r17/wclub, clear
. matrix dissimilarity clubD = , variables Jaccard dissim(oneminus)
. clustermat waverage clubD, name(clubwav) clear labelvar(question)
. cluster dendrogram clubwav, labels(question)
> title(Weighted-average linkage clustering)
> xlabel(, angle(90) labsize(*.75)) ytitle(1 - Jaccard similarity, suffix)
```



30 名の女性の回答から、biog（伝記を読んで楽しむ）と hist（歴史を読んで楽しむ）という 2 つの質問は最も関連性が高いことがわかります。bird（鳥を飼っている）という質問は残りの質問のグループと最後に結合されているため他の質問と最も関連が低いと考えられます。