Craig Van Vliet & Ryan Thomas

ECO 7100 – Econometrics 1

April 6th, 2019

# Lab: Hierarchical Clustering

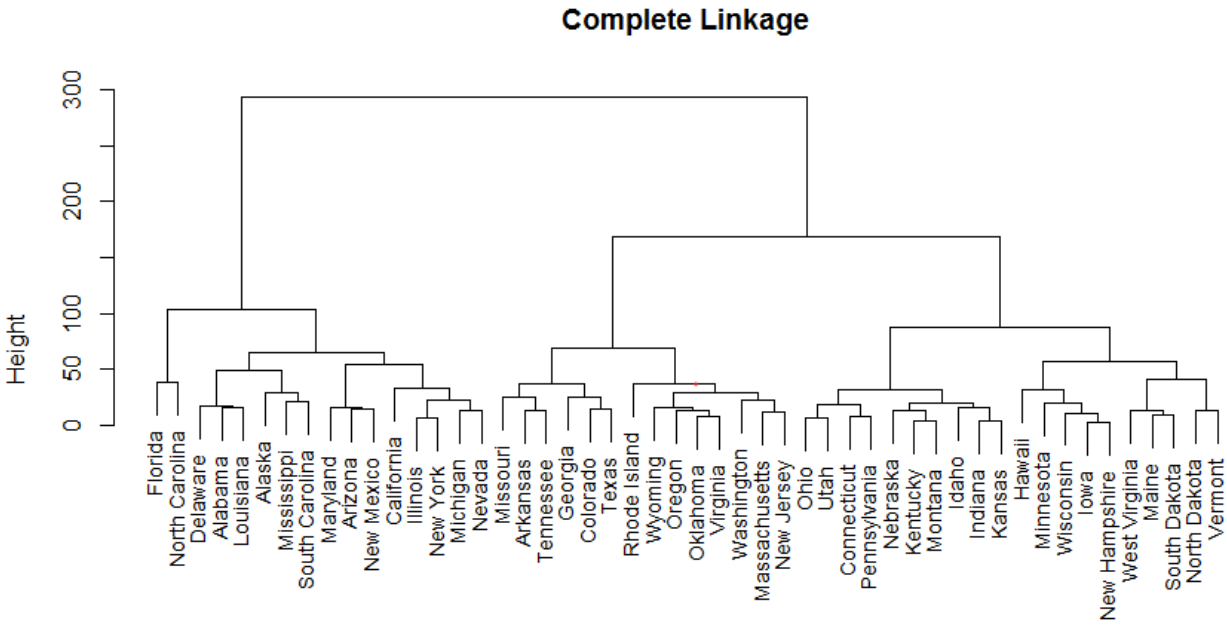
Assignment Submission

1) Consider the USArrests data. We will now perform hierarchical clustering on the states.

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

```
library(ISLR)
set.seed(2)
summary(USArrests)

hc.complete=hclust(dist(USArrests,method="euclidian"), method="complete")
plot(hc.complete,main="Complete Linkage", xlab="", sub="", cex=.9)
```



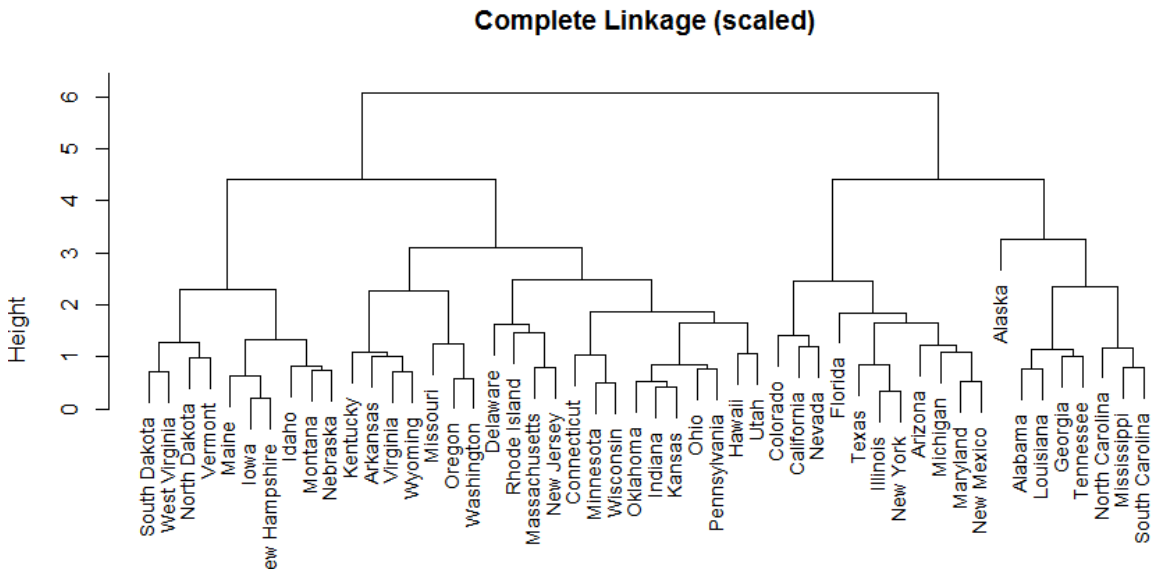

(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

```
cut=cutree(hc.complete, 3)
cutdat=cbind(USArrests,cut)
rownames(cutdat[cut==1,])
rownames(cutdat[cut==2,])
rownames(cutdat[cut==3,])
```

```
> rownames(cutdat[cut==1,])
 [1] "Alabama"        "Alaska"         "Arizona"        "California"
 [5] "Delaware"       "Florida"        "Illinois"       "Louisiana"
 [9] "Maryland"       "Michigan"       "Mississippi"    "Nevada"
[13] "New Mexico"     "New York"       "North Carolina" "South Carolina"
> rownames(cutdat[cut==2,])
 [1] "Arkansas"       "Colorado"       "Georgia"        "Massachusetts"
 [5] "Missouri"       "New Jersey"     "Oklahoma"       "Oregon"
 [9] "Rhode Island"   "Tennessee"      "Texas"          "Virginia"
[13] "Washington"     "Wyoming"
> rownames(cutdat[cut==3,])
 [1] "Connecticut"    "Hawaii"         "Idaho"          "Indiana"
 [5] "Iowa"           "Kansas"         "Kentucky"       "Maine"
 [9] "Minnesota"      "Montana"        "Nebraska"       "New Hampshire"
[13] "North Dakota"   "Ohio"           "Pennsylvania"   "South Dakota"
[17] "Utah"           "Vermont"        "West Virginia"  "Wisconsin"
```

(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

```
xsc=scale(USArrests)
hc.scaled=hclust(dist(xsc,method="euclidian"), method="complete")
plot(hc.scaled,main="Complete Linkage (scaled)", xlab="", sub="", cex=.9)
```



```
cut2=cutree(hc.scaled, 3)
cutdat2=cbind(USArrests,cut2)
rownames(cutdat2[cut2==1,])
rownames(cutdat2[cut2==2,])
rownames(cutdat2[cut2==3,])
```

```
> rownames(cutdat2[cut2==1,])
[1] "Alabama"        "Alaska"         "Georgia"        "Louisiana"
[5] "Mississippi"    "North Carolina" "South Carolina" "Tennessee"
> rownames(cutdat2[cut2==2,])
 [1] "Arizona"    "California" "Colorado"   "Florida"    "Illinois"
 [6] "Maryland"   "Michigan"   "Nevada"     "New Mexico" "New York"
[11] "Texas"
> rownames(cutdat2[cut2==3,])
 [1] "Arkansas"      "Connecticut"   "Delaware"      "Hawaii"
 [5] "Idaho"         "Indiana"       "Iowa"          "Kansas"
 [9] "Kentucky"      "Maine"         "Massachusetts" "Minnesota"
[13] "Missouri"      "Montana"       "Nebraska"      "New Hampshire"
[17] "New Jersey"    "North Dakota"  "Ohio"          "Oklahoma"
[21] "Oregon"        "Pennsylvania"  "Rhode Island"  "South Dakota"
[25] "Utah"          "Vermont"       "Virginia"      "Washington"
[29] "West Virginia" "Wisconsin"     "Wyoming"
```

(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

Scaling the variables ensures that no single variable has an unfairly strong influence on clustering. This is especially useful if variables are measured in different units or have widely ranging variances, especially if clustering using Euclidian distance. With this data, the variable "Assault" has a much higher variance than other variables and therefore has a greater influence on clustering. States end up being clustered according to Assault rates more than according to any other variable if the data is not scaled.

```
> var(xsc)
             Murder   Assault   UrbanPop      Rape
Murder   1.00000000 0.8018733 0.06957262 0.5635788
Assault  0.80187331 1.0000000 0.25887170 0.6652412
UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
Rape     0.56357883 0.6652412 0.41134124 1.0000000
> var(USArrests)
             Murder   Assault   UrbanPop      Rape
Murder    18.970465  291.0624   4.386204  22.99141
Assault  291.062367 6945.1657 312.275102 519.26906
UrbanPop   4.386204  312.2751 209.518776  55.76808
Rape      22.991412  519.2691  55.768082  87.72916
```

We are able to see the benefits of clustering by analyzing the differences in range and mean values of the different variables between clusters. First, we will print a summary of these measures for the un-scaled data.

**US Arrest data, clustered (not scaled)**

```
> summary(cutdat[cut==1,])
     Murder          Assault         UrbanPop          Rape           cut
 Min.   : 5.90   Min.   :236.0   Min.   :44.00   Min.   :15.80   Min.   :1
 1st Qu.:10.30   1st Qu.:251.2   1st Qu.:55.50   1st Qu.:21.95   1st Qu.:1
 Median :11.75   Median :261.0   Median :71.00   Median :26.95   Median :1
 Mean   :11.81   Mean   :272.6   Mean   :68.31   Mean   :28.38   Mean   :1
 3rd Qu.:13.50   3rd Qu.:287.2   3rd Qu.:80.25   3rd Qu.:32.85   3rd Qu.:1
 Max.   :16.10   Max.   :337.0   Max.   :91.00   Max.   :46.00   Max.   :1
> summary(cutdat[cut==2,])
     Murder          Assault         UrbanPop          Rape           cut
 Min.   : 3.400   Min.   :145.0   Min.   :50.00   Min.   : 8.30   Min.   :2
 1st Qu.: 5.325   1st Qu.:156.8   1st Qu.:60.75   1st Qu.:18.98   1st Qu.:2
 Median : 7.650   Median :167.5   Median :69.00   Median :23.10   Median :2
 Mean   : 8.214   Mean   :173.3   Mean   :70.64   Mean   :22.84   Mean   :2
 3rd Qu.: 8.950   3rd Qu.:189.5   3rd Qu.:79.50   3rd Qu.:26.73   3rd Qu.:2
 Max.   :17.400   Max.   :211.0   Max.   :89.00   Max.   :38.70   Max.   :2
> summary(cutdat[cut==3,])
     Murder          Assault         UrbanPop          Rape           cut
 Min.   :0.80   Min.   : 45.00   Min.   :32.00   Min.   : 7.30   Min.   :3
 1st Qu.:2.50   1st Qu.: 56.75   1st Qu.:51.75   1st Qu.:11.03   1st Qu.:3
 Median :3.55   Median : 94.00   Median :59.50   Median :14.55   Median :3
 Mean   :4.27   Mean   : 87.55   Mean   :59.75   Mean   :14.39   Mean   :3
 3rd Qu.:6.00   3rd Qu.:110.75   3rd Qu.:67.50   3rd Qu.:16.88   3rd Qu.:3
 Max.   :9.70   Max.   :120.00   Max.   :83.00   Max.   :22.90   Max.   :3
```

Without scaling, States are clustered together mainly based on Assault rate. The cluster ranges for the Assault measure do not overlap, while the ranges of all other measures have significant overlap between clusters.

**US Arrest data, clustered (scaled)**

```
> summary(cutdat2[cut2==1,])
     Murder          Assault         UrbanPop          Rape           cut2
 Min.   :10.00   Min.   :188.0   Min.   :44.00   Min.   :16.10   Min.   :1
 1st Qu.:13.15   1st Qu.:229.8   1st Qu.:47.25   1st Qu.:20.18   1st Qu.:1
 Median :13.80   Median :254.0   Median :53.00   Median :22.35   Median :1
 Mean   :14.09   Mean   :252.8   Mean   :53.50   Mean   :24.54   Mean   :1
 3rd Qu.:15.57   3rd Qu.:267.0   3rd Qu.:59.25   3rd Qu.:26.07   3rd Qu.:1
 Max.   :17.40   Max.   :337.0   Max.   :66.00   Max.   :44.50   Max.   :1
> summary(cutdat2[cut2==2,])
     Murder          Assault         UrbanPop          Rape           cut2
 Min.   : 7.90   Min.   :201.0   Min.   :67.00   Min.   :24.00   Min.   :2
 1st Qu.: 9.70   1st Qu.:250.5   1st Qu.:76.00   1st Qu.:26.95   1st Qu.:2
 Median :11.30   Median :255.0   Median :80.00   Median :31.90   Median :2
 Mean   :11.05   Mean   :264.1   Mean   :79.09   Mean   :32.62   Mean   :2
 3rd Qu.:12.15   3rd Qu.:289.5   3rd Qu.:82.00   3rd Qu.:36.90   3rd Qu.:2
 Max.   :15.40   Max.   :335.0   Max.   :91.00   Max.   :46.00   Max.   :2
> summary(cutdat2[cut2==3,])
     Murder          Assault         UrbanPop          Rape           cut2
 Min.   :0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30   Min.   :3
 1st Qu.:2.950   1st Qu.: 82.0   1st Qu.:53.50   1st Qu.:11.25   1st Qu.:3
 Median :4.900   Median :113.0   Median :66.00   Median :16.30   Median :3
 Mean   :5.003   Mean   :116.5   Mean   :63.84   Mean   :16.34   Mean   :3
 3rd Qu.:6.700   3rd Qu.:153.5   3rd Qu.:72.50   3rd Qu.:20.10   3rd Qu.:3
 Max.   :9.700   Max.   :238.0   Max.   :89.00   Max.   :29.30   Max.   :3
```

After scaling, states are clustered with less range overlap. Cluster 1 seems to represent states with the highest rate of murder charges. Cluster 2 represents states with the highest urban population and rape charges. Cluster 3 is the largest group and it represents the states with relatively lower crime statistics. Scaling made the clusters more interpretable and focused.