chapter9

UDOH DANIEL

9/25/2022

CHAPTER 9

```
                     STATISTICS FOR THREE OR MORE VARIABLES
           The first episode was about computing a multiple regression. The multiple regression is the most common and powerf
ul technique used were several variables are used collectively to predict scores on a single outcome variable or qualitative
outcome. A basic multiple regression can be created and we can look for the summary of the regression created. More detailed
summaries can also be gotten with their functions which are: anova(), coef(), confint(), resid(), hist(). There is a possibi
lity of stepwise variables selection (forward and backward).
              The second episode was about comparing means with a two-factor ANOVA. It allows us to use two categorical pred
ictive variable and a quantitative outcome. A simple boxplot can be created for dataset with more than one variable at the s
ame time, (e. g) boxplot (breaks~woool*tension, data=warpbreaks). For additional informations on model we can use the functi
on model.tables(). A post-hoc test can be done, (e. g) TukeyHSD().
                The third episode was about conducting a cluster analysis which is the ability to group cases based on simil
arities and scores on the variables in a dataset. There are three major categories of clustering which includes: split into
set number of clusters (e.g) kmeans, hierarchical (start separate and combine) and dividing (start with a single group and s
plit). To do a hierarchical clustering, we need a distance matrix/dissimilarity matrix. To use distance matrix for clusterin
g, the function hclust() is used. We can plot a dendogram (branches) of clusters using plot() function. The function cutree
() is used to specify the heights and groups of the plot. It can be done for a single groups or several groups. To draw boxe
s around clusters we use the function rect.hclust(). To do a k-means clustering we use the function kmeans(). To make a grap
h based on k-means we use the function clusplot().
               The fourth episode was about conducting a principal components/factor analysis. The primary empirical differe
nce between a components and a factor model is the treatment of the variances for each items. The functions prcomp() or prin
comp[() are used for conducting a principal components model. We can make a screeenplot with the function plot(). To see how
cases load we use the function product(). We can create a biplot using the function biplot(). For performing a factor analys
is, we use the function factanal() to check for chi-square and p-values.
```

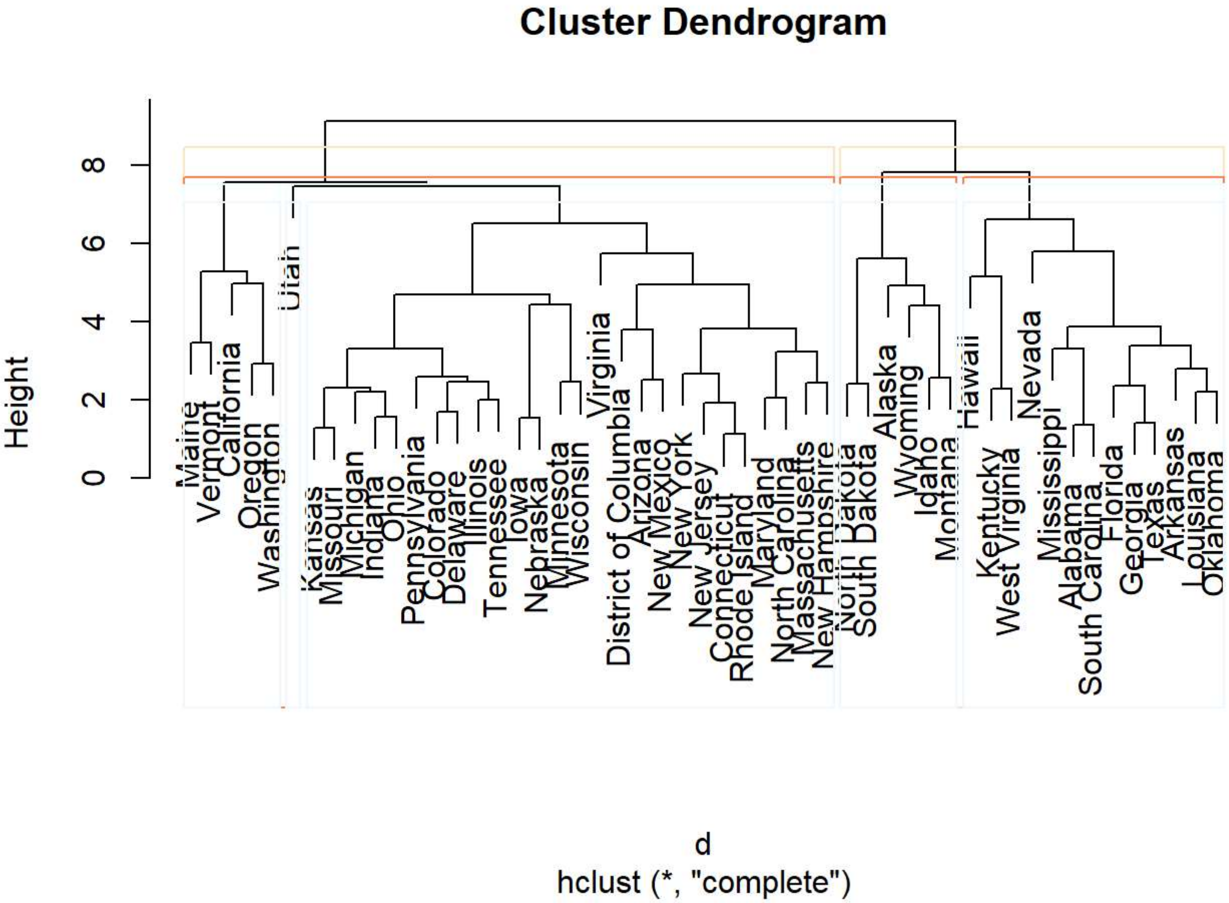
###load dataset

```
scd <-read.csv("C:/Users/OLAJIDE/Videos/R Statistics Essential Training/Exercise Files/Ch09/09_05_Challenge/StateClusterData.csv",header=TRUE)
rownames(scd)<-scd[,1]
scd[,1]<-NULL
```

###hierarchical cluster

```
d <- dist(scd)
c <- hclust(d)
```

```
plot(c)
###draw boxes around clusters
rect.hclust(c, k = 2, border = "bisque")
rect.hclust(c, k = 3, border = "coral")
rect.hclust(c, k = 4, border = "azure")
rect.hclust(c, k = 5, border = "aliceblue")
```


Cluster Dendrogram
Height
0
2
4
6
8
Maine
Vermont
California
Oregon
Washington
Utah
Kansas
Missouri
Michigan
Indiana
Ohio
Pennsylvania
Colorado
Delaware
Illinois
Tennessee
Iowa
Nebraska
Minnesota
Wisconsin
Virginia
District of Columbia
Arizona
New Mexico
New York
New Jersey
Connecticut
Rhode Island
Maryland
North Carolina
Massachusetts
New Hampshire
North Dakota
South Dakota
Alaska
Wyoming
Idaho
Montana
Hawaii
Kentucky
West Virginia
Nevada
Mississippi
Alabama
South Carolina
Florida
Georgia
Texas
Arkansas
Louisiana
Oklahoma
d
hclust (*, "complete")