Ex1)

1. Check the multivariate gaussianity assumption:

Pvalue of the Shapiro test is 0.664 -> we accept the hyp of gaussianity

So the vector with components: the number of accesses to the online store, the number of purchases of men's clothing and the number of purchases of women's clothing is normally distributed.

CR for the mean (ellipsoidal region)

CR= { m \in R^3 t.c. n \* (x.mean-m)' %\*% x.invcov %\*% (x.mean-m) < cfr.fisher }

Center :

accesses men women

211.12500 20.79167 27.04167

Direction of the axis:

[1,] 0.98529663 0.05642585 0.1612658

[2,] 0.09957569 -0.95665950 -0.2736554

[3,] 0.13883524 0.28568993 -0.9482120

Length of the semi-axes:

15.693925 5.842251 3.965390

Issues? Non penso ci siano

1. T2- simultaneous CI

Inf center sup

accesses 195.64510 211.12500 226.60490

men 14.88768 20.79167 26.69566

women 22.38644 27.04167 31.69689

men+women 40.56411 47.83333 55.10255

1. Test:

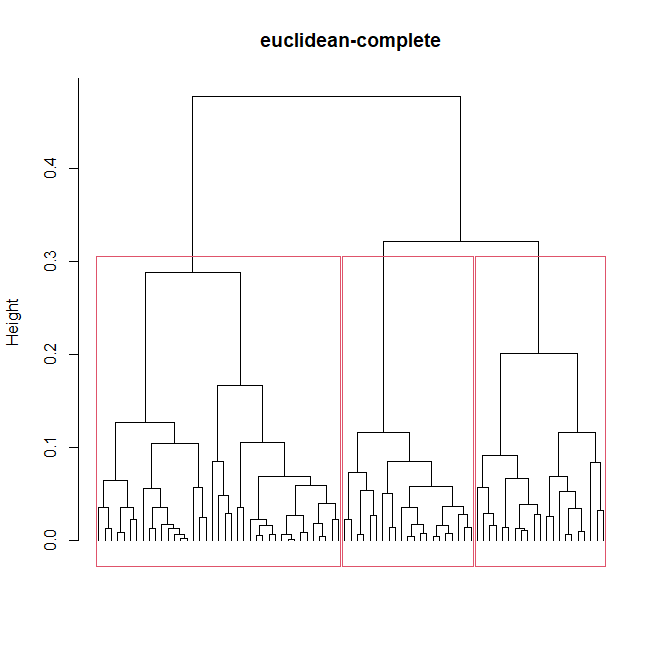
H1 : men+women > 0.2 \*access

-0.2 acess + men + women > 0 -> metto in h1 perchè è quello che voglio dimostrare

Pvalue = 0.005304853 -> reject so we have proved our hypothesis

Ex2)

1. Dendogram



Data within each cluster:

Ni = numerosita cluster i

N1 = 39, N2 = 21, N3 = 21

Mean within each cluster:

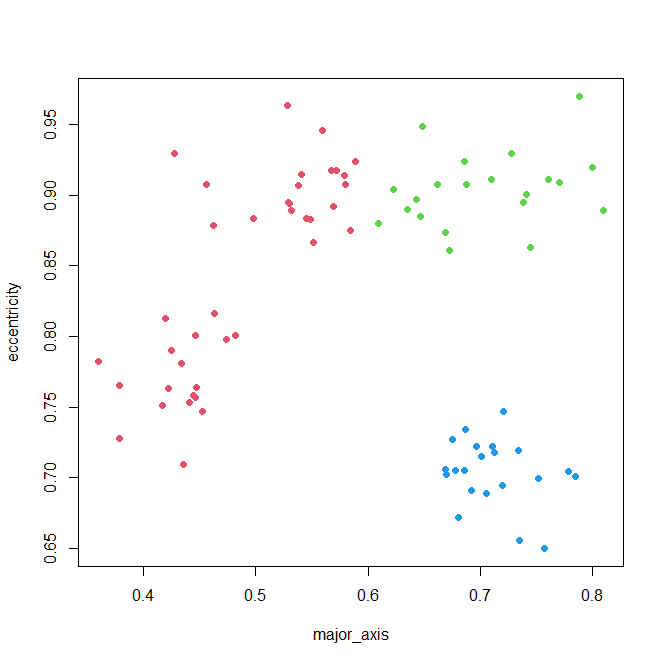
clust1: 0.4884103 0.8428462

clust2: 0.7037143 0.9037619

clust3: 0.7118571 0.7036667

Come funzione meglio usare colMeans: sono sicura che fa la media per colonne!

Plot:

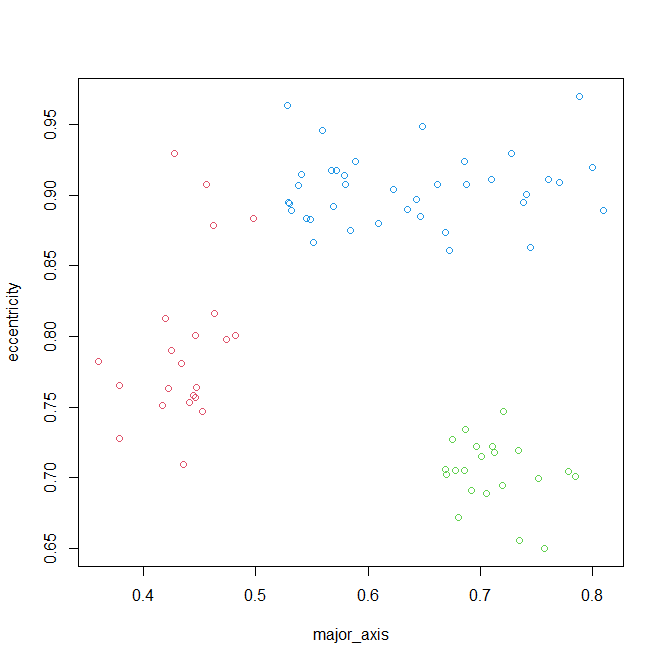


1. To evaluate the classifier we can compute the cophenetic coeff which measures the correlation between the original distance matrix and the new distance matrix, so an high correlation means that the dendogram is capturing the cluster structure

We have a value of 0.78 which is quite good

Anyway a possible issue that we can observe from the dendogram is that the clustering may be not much robust. And also from the plot of the clusters maybe we would expect that the top part of the red cluster should belong to the green one.

Try with k-means



There are still some points classified as red and maybe we would expect blu but it seems better.

Means:

1: 0.4366364 0.7944545

2: 0.7118571 0.7036667

3: 0.6373684 0.9045263

Numerosity in each cluster:

N1 = 22 , N2= 21, N3= 38

We proceed with the clustering given by k-means.

ATT: un’altra idea è continuare con hierarchical cambiando però il linkage, qui vediamo dal plot che i 3 cluster sono abbastanza separati quindi ci aspettiamo che single linkage vada bene

Infatti se rifacciamo con linkage otteniamo i cluster che ci saremmo aspettati ma coph si abbassa 0.65

1. Bonf:

Prima pensavo di essere nel frame di ANOVA ma invece penso sia solo inferenza su parametri di 3 distribuzioni normali quindi

Create the 3 dataset with the data of each cluster and check the normality inside each cluster, with the Shapiro test we ca see that the hyp of gaussianity is met (gaussianity fatta con Shapiro visto che è univariato).

Now we can build , within each cluster, the bonf intervals for mean and variance (quindi abbiamo k=2 e lo stiamo facendo separatamente in ogni cluster, pensiamo che considerare k=6 non abbia senso visto che sono 3 dataset diversi)

Cluster1:

* Mean

inf center sup

major\_axis 0.4637301 0.4884103 0.5130904

* Variance

inf center sup

0.002755329 0.004361775 0.007823855

Stessi calcoli per gli altri cluster

Ex3)

1. y = beta\_0 + beta\_1 \* z1 + beta\_2 \* z2 + beta\_3 \* z3 + beta\_4 \* z4+ Eps

where z1,.., z4 are the regressors and y is the target (rate)

z1 = rain ; z2 = hardness, z3 = coarse, z4 = fine

Assumptions for the model: E(Eps) = 0 and Var(Eps) = sigma^2

Estimates parameters:

Beta\_0 Beta\_1 Beta\_2 Beta\_3 Beta\_4

20.514647296 0.006115268 0.011423423 0.385216816 0.195448455

1. remove z2 since it does not seem significant indeed in the output of the model we get a pvalue around 81%

The model becomes:

y = beta\_0 + beta\_1 \* z1 + beta\_2\* z3 + beta\_3 \* z4+ Eps

Parameters:

Beta\_0 Beta\_1 Beta\_2 Beta\_3

20.560257374 0.006131644 0.386020255 0.194141815

1. Construct a linear hypothesis test based on the last model

H0 : beta\_2 – 2\* beta3 = 0

H1 : != 0

Pvalue = 0.9836 -> assume H0

Constrained model :

(All’inizio avevo definito z5 = z3-2z4 ma non viene significativa (e ha senso perché so che l’impatto della combinazione è nullo))

Quindi l’idea è procedere sfrittando la relazione che abbiamo trovato

beta\_2 = 2\*beta\_3

y = beta\_0 + beta\_1 \* z1 + beta\_2\* z3 + beta\_3 \* z4+ Eps =

beta\_0 + beta\_1 \* z1 + beta\_2\* z3 + 0.5\*beta\_2 \* z4+ Eps =

beta\_0 + beta\_1 \* z1 + beta\_2\* (z3 + 0.5\*z4) + Eps

define z5 = z3 + 0.5\*z4 and fit the model

y = beta\_0 + beta\_1 \* z1 + beta\_2\* z5+ Eps

parameters

Beta0 beta 1 beta 2

22.117461901 0.005921752 0.419340895 (sono della vecchia sol)

1. Pointwise estiamate = 30.4561

CI = [30.24241, 30.66978]

Ex4)

1. We are assuming stationarity since the mean in the whole space is the same (we don’t have any regressor)

# Hyp: second-order stationarity and isotropy

Estimated a0 = 263.5391 -> è questo? Sì

Estimated parameters for the residuals:

model psill range

1 Nug 0.000 0.0000

2 Sph 4659.543 386.7413 -> questo? Sì

Modello stimato per delta:

spherical model with:

* Nugget = 0
* Range = 0 + 386.7413
* Sill = 0 + 4659.543

1. Now we don’t assume stationarity and insert two regressors in our model: the categorical variable which accounts for winter/ no winter and the numerical variable which translate the distance

ATT: ricordati di meyttere anche l’interazione dummy:distance nel modello altrimenti non otterrei a1,g

v2 <- variogram( resp ~ dummy\_no + distance + dummy\_no:distance, dataset)

model psill range

1 Nug 305.1408 0.0000

2 Sph 646.3564 526.2371

Parameters:

BHO, noi l’abbiamo fatto solo con dummy

1. Choose model b?
2. The model suggests a price of 386.0059 per nights -> 4\* 386.0059 = 1544.024, overall