Ex1)

1. Since the variances of the original variables are very different with some of them meaningfully larger than the others we choose to standardize the variables in order to have comparable variance for all the variables.

Performing PCA: loadings of PC1:

Energy\_kcal 0.368

Protein\_g -0.538

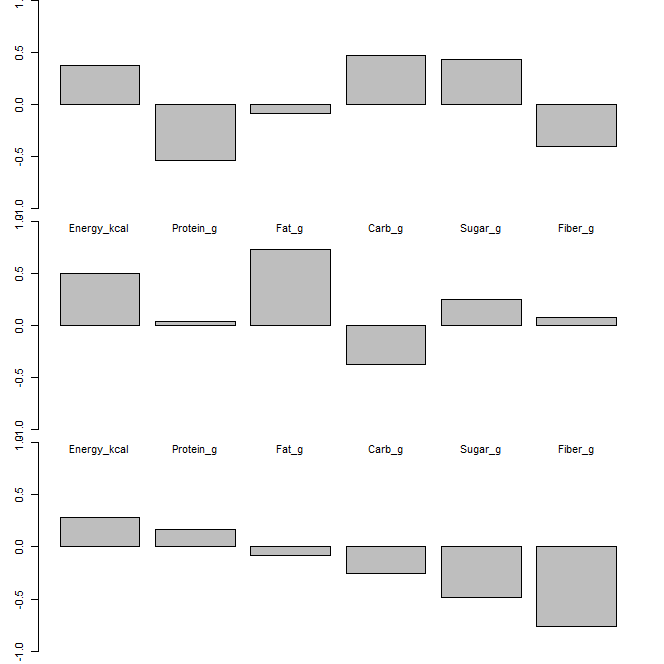
Fat\_g 0

Carb\_g 0.466

Sugar\_g 0.434

Fiber\_g -0.400

1. Loadings:



Interpretation:

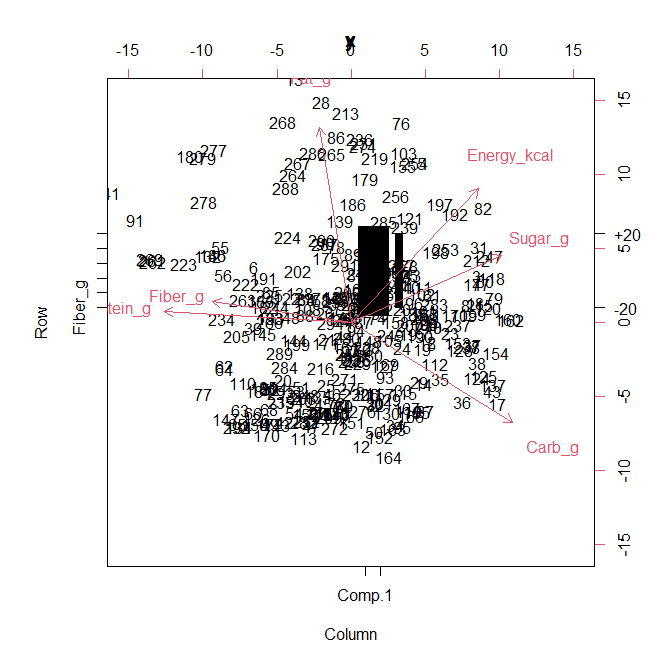
PC1 shows a contrast between protein and fiber against energy, carb and sugar.

Pc2 is again a contrast mainly between energy and fat against carb

PC3 is the (opposite-> NO ) of a weighted mean between sugar and fiber (we consider only this variables since the loadings of the others are significantly lower than these 2)

* Se PC3 negativa allora ho alta media, invece se è positiva bassa

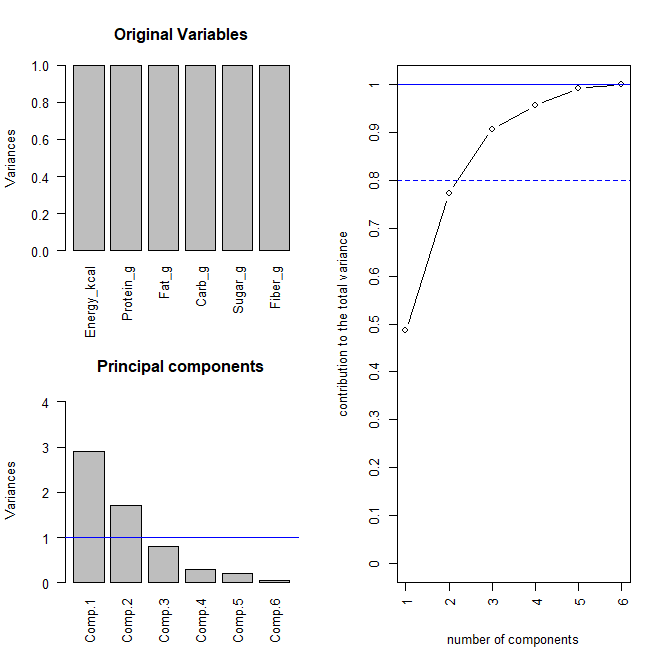
1. Biplot:



Cereals with positive scores both for PC1 and PC2 are characterized by an higher value of energy and sugar and instead a lower value for figer and protein

Rivedi interpretazione

1. Plot



From the plot of the explained variability we can see that choosing 3 PC we explain around 90% of the variability which is a good amount and moreover adding others PC the gain in terms of explained variability is little so it is not worth it -> we choose 3 PC

Variance explained by each PC:

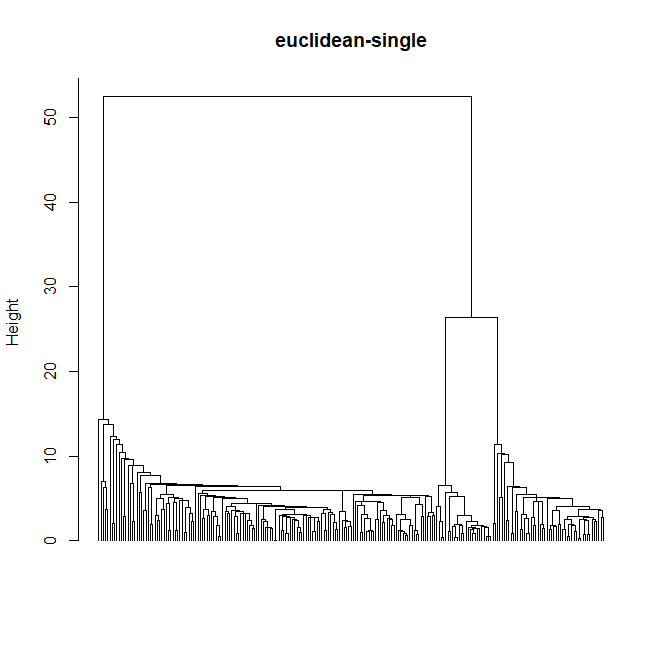
0.486122577 0.286261161 0.134653372

Cumulative proportion of explained variance with the first 3 Pc = 0.9070371

1. Come faccio la proiezione di solo un dato?

Ex2)

1. The obtain the dendogram:



So the appropriate number of clusters seems to be k=3 so that we have 3 clusters and the choice seems to be robust.

1. Centroids of the clusters:

minutes artists

clust1 199.9639 29.5506

clust2 59.9540 44.8832

clust3 48.0832 3.7916

size:

clust 1 = 150 ; clust 2 = 50 ; clust3 = 25

cophenetic coeff = 0.8953819

1. Under the assumptions that each of the 3 populations identified by the clustering is a bivariate gaussian we can compute the bonf intervals for the components of the mean in each cluster at global level 95% with k=p=2 since we perform a bonf correction in each cluster:

* Clust 1

inf center sup

minutes 194.36499 199.9639 205.56274

artists 27.59271 29.5506 31.50849

* Clust 2

inf center sup

minutes 56.58700 59.9540 63.32100

artists 42.40518 44.8832 47.36122

* Clust 3

inf center sup

minutes 43.742873 48.0832 52.423527

artists 3.193882 3.7916 4.389318

Ex3)

1. Parameters:

* Beta

Beta0 beta1

red -1.2533849 0.7272803

white -0.7004341 0.5962484

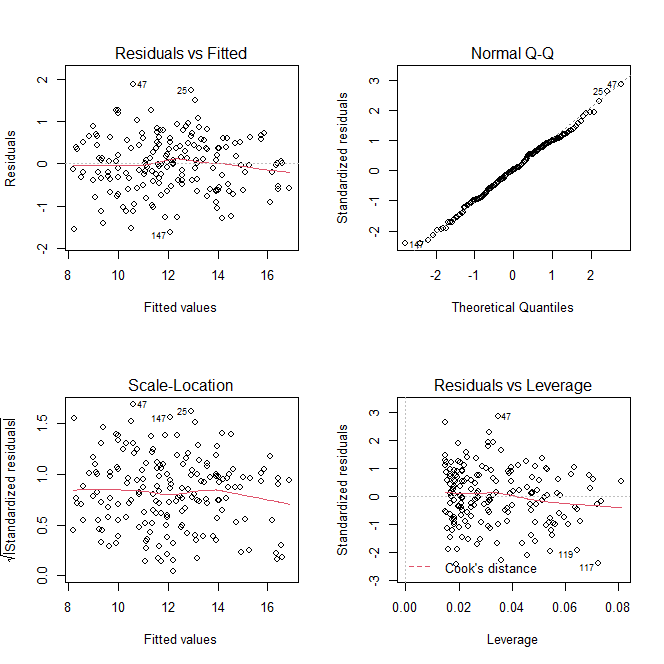
rose -1.7119121 0.6621416

* Sigma = 0.6726

Assumption : Eps ~ N(0, sigma^2) -> mean around 0, homoschedastic and gaussian

Where the gaussianity is needed to make inference on the paameters.

* Check that the residuals are centered in 0 and homogeneous:



We can observe from the plot of the residuals against the fitted values that they have mean around 0 and thery are quite homoschedastic.

From the qqplot we see that the gaussianity assumption seem to be met and performing a Shapiro test this is confirmed since we obtain a pvalue of 0.7382 -> accept gaussianity.

1. Perform a linear hypothesis test checking if all the cieff related to the dummy variables (we have 2 dummies one which encode the type red and the other type white) can be put together at 0

So In H0 we have that they all are = 0

The pvalue of the test is the numerical 0 so we reject H0 and we can state that there is significant dependence of the mean of alcohol on the type of wine

* Same test but in H0 now we have that all the ceff related with sugar are = 0:

Again we obtain as pvalue the numerical 0 so we reject and we state that there is significant dependence of the mean of alcohol on the sugar content

1. Performing a test to check if we can remove the dummies without interaction we obtain a pvalue of 0.54 so we can remove them keeping only the interactions of the dummies with sugar, so we fit the model:

alcohol ~ sugar + sugar:dummy\_red + sugar:dummy\_white

and now each regressor is significant so we can estimate the parameters

* Beta

Beta0 beta1

red -1.243293 0.7267769

white -1.243293 0.6223343

rose -1.243293 0.6392491

so all groups have the same intercept but they have different slope

* Sigma = 0.6711

1. Prediction interval 11.53151 15.05298

Pointwise prediction = 13.29224

Ex4)

1. We fit a non-stationary model since we want to include the regressor given by the logarithm of the chlorofill . So we are not under the assumption of stationarity but we are assuming isotropy.

Model estimated for delta:

model psill range

Exp 0.6383447 71791.88

* Exponential model without nugget, with sill = 0.6383447 and range = 71791.88

Parameters of the model:

* A0 = 2.745206
* A1 = 0.739114

Assumptions: isotropy e stazionarieta diìei residui

1. Fit a stationary model where log.chlorofill is the response.

The estimated model has the following parameters:

model psill range

Sph 3.638485 38953.76

* Spherical model with sill = 3.638485 and range = 38953.76

The pointwise estimate for the log.chlorofill in the new location is 3.983721.

Using this value to predict the log.sights we find as prediction log.sights(s0)= 5.739903

Nome-col-resp è log.sights? quindi la aggiungo al dataset? sì

1. The variance of the prediction of log.sights at point b is 0.4041272.

I think that this variance is not fully representative of the uncertainty because we have obtained the prediction using as regressor another point prediction which therefore is also affected by uncertainty.

The variance reported here does not account for the uncertainty of the prediction of the regressor so the global variance will be higher.