Ex1)

1. Assumptions of the model:

* Normality in each of the 6 subrgoups given by the combination of the 2 groupings location and type:

Pvalue = 0.9418255 0.6449031 0.2579667 0.3074365 0.2909484 0.6741621

These are the pvalue of the shapirro test in each subgroups -> assume normality

* Homogeneity: we perform the barlett test to check if we can assume same variance in each group

Pvalue = 0.2473 -> assume H0 which is the hypothesis of homogeneity

* The assumptions are verified

Complete anova model:

### Model with interaction (complete model):

### x.ijk = mu + tau.i + beta.j + gamma.ij + eps.ijk; eps.ijk~N(0,sigma^2),

### i=1,2 (effect location), j=1,2 (effecttype)

1. From the summary of the model we can read the pvalue of the test which tests the significant of the interaction between location and type:

H0: gamma.ij = 0 per ogni i,j vs H1: (H0)^c

i.e.,

H0: There is no significant interaction between the factors

H1: There exists a significant interaction between the factors

In this case the pvalue is 0.389 -> accept H0

Proceed with an additive model removing the interaction:

### x.ijk = mu + tau.i + beta.j + eps.ijk; eps.ijk~N(0,sigma^2),

### i=1,2 (effect location), j=1,2 (effect type)

Again from the output we read the pvalue of the test which test the significance of each grouping and in particular group1 (location) does not seem to be significant:

### H0: tau.1 = tau.2 = 0 vs H1: (H0)^c

### i.e.,

### H0: The effect group1 doesn't significantly influence the price

### H1: The effect group1 significantly influences the price

Indeed the pvalue is 0.719

Reduce the model removing the location:

Modello finale : x.jk = mu + beta.j + eps.jk; eps.jk~N(0,sigma^2)

1. Estimate of sigma^2 = 19955.57

Estimate for mu = 872.5583

Estimates for betaj : -> scrivi quali sono i “ j “ !

beta1 beta2 beta

120.46667 -40.33333 -80.13333

1. "apt – bb” = -116.52284 36.92284

"apt - hotel" = -277.3228 -123.8772

"bb - hotel" =-237.52284 -84.07716

We see that the Bonf intervals for the difference of the mean of the price between apt and hotel and the difference between bb and hotel do not contain 0 which means that the 2 means are significantly different instead from these intervals it does not seem to be a difference between the mean price in apt and the one in bb. So we can say that the type is significant due to the presence of hotel indeed the price oh hotels is larger than the one in apt and bb (indeed the intervals for the difference is negative)

BF spiega quali sono i livelli di type responsabili della significatività del tipo

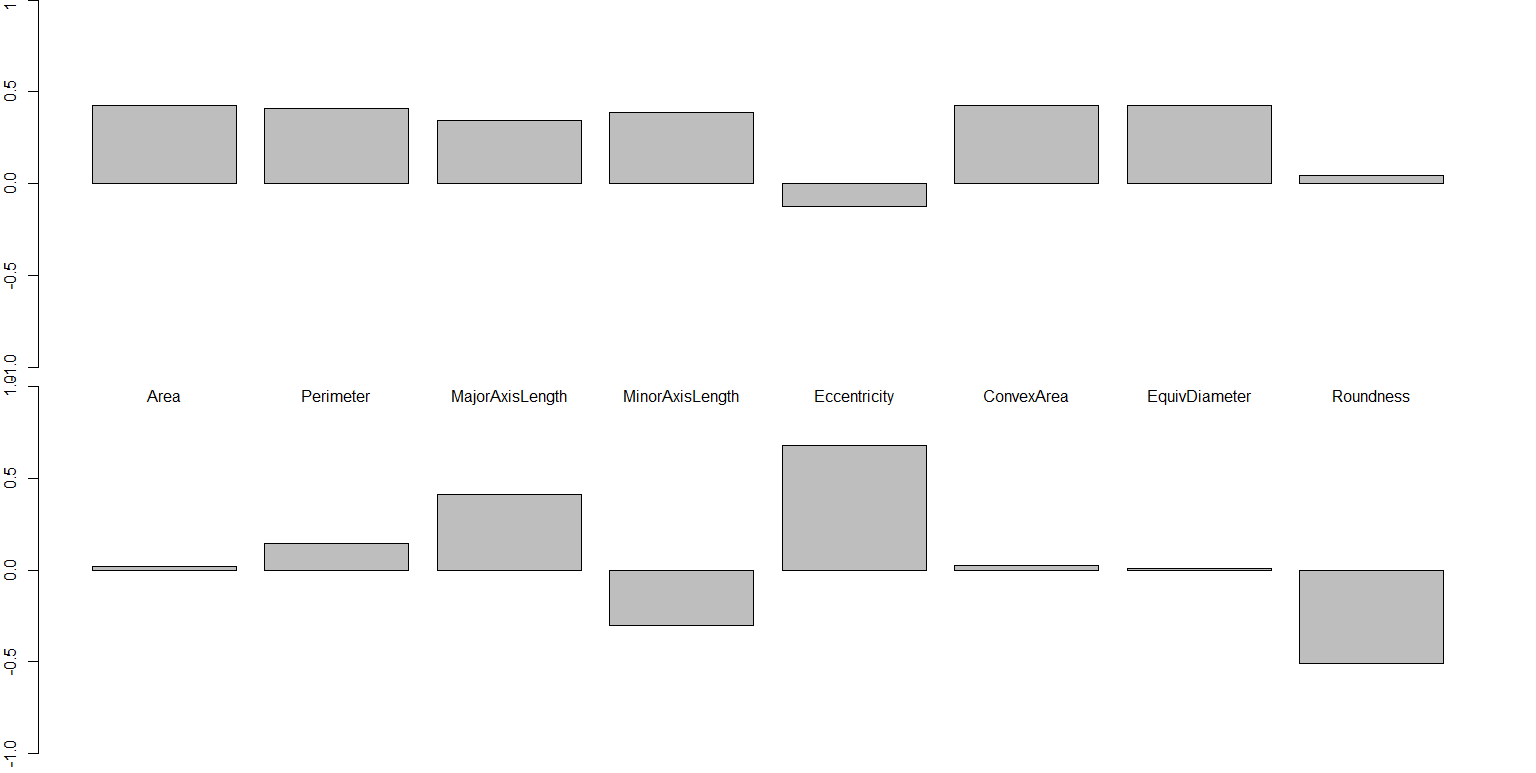
Ex2)

1. Observing the boxplot of the numerical variables we see that the 2 related with the area have a much larger variability than the others so we scale and center the data and perform pca on the scaled data.

Performing PCA we see that the first 2 PC explain almost the 90% of the variability and adding also the third PC we reach 99% so we could keep 2 or 3 PC

* Visto che l’interpretazione della terza non è super chiara magari preferiremmo tenerne 2

1. First 2 Pc:



The first PC is a weighted mean of all the variables except for eccentricity and roundness whose weights are almost 0. (coherently with the boxplot since those 2 variables are the ones with smaller variance)

The second PC shows a contrast between the perimeter (questa si può togliere), the majoraxis length and the eccentricity against the minor axis length and the roundness.

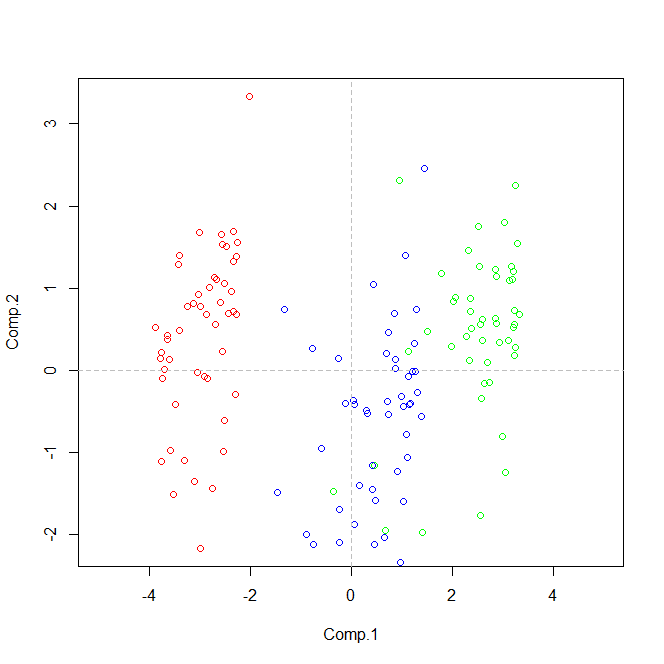
Plotting the scatterplot based on the type:

Plot in green the data associated with the adzuky type and so we see that they are characterize by a positive value of PC1, which means that they have a large weighted mean

Instead in red we see the ones of type cannellini which are characterized by negative values of PC1 so they have a smaller weighted mean

Blue point have a standard size since Pc1 is almost 0 -> valori intorno alla media

NB: gli scores sono centrati nella media



1. Build a dataframe with only the scores of PC1 and Pc2 associated with the type cannellini. In order to build a CR we need to test the assumption of bivariate normality so we perform a mcshapuro test obtaining a pvalue of 47% -> assume gaussianity.

CR for the mean:

CR = { m \in R^2 t.c. n \* (x.mean-m)' %\*% x.invcov %\*% (x.mean-m) < cfr.fisher }

Where x is the vector in R2 which contains our data.

The center, which is given by the mean, is:

Comp.1 Comp.2

-2.9513269 0.4281307

The direction of the axes are:

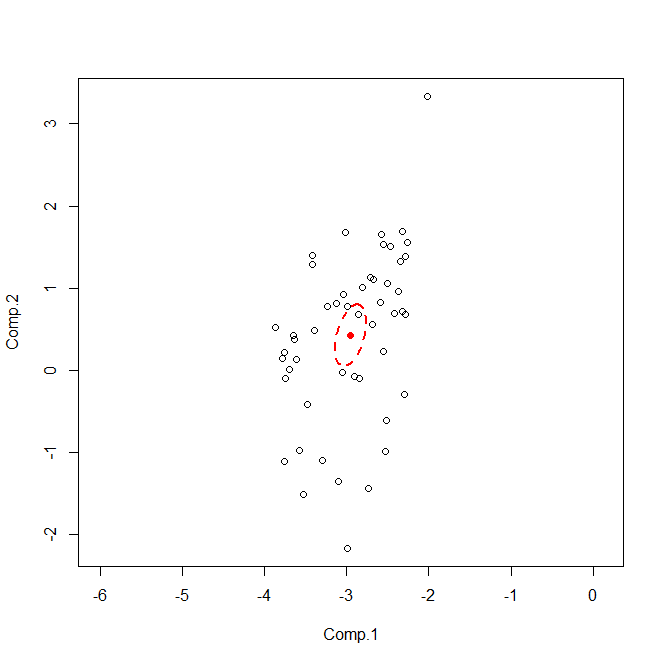
0.2570354, -0.9664020

0.9664020, 0.2570354

The length of the semiaxis are:

0.3569484, 0.1546133

Plot:



Piccola? Forse ok perché sono molto concentrati quindi non ho punti lontani che me la fanno espandere -> pensala così: CR piccola mi sto sbilanciando molto nel dire che è lì dentro e posso farlo perché i dati sono belli

Ex3)

1. We create a dummy for the sex and fit the linear model with all the covariates in the dataset

y = b0 + b1\*age + b2\*height + b3\*dummy\_sex + b4\*distance + b5\*siblings + b6\*computertime + b7\*exercisehours + b8\*musiccds + b9\*playgames

Parameters:

Beta:

(Intercept) age height dummy\_sex distance

6.8710519473 0.1948575363 -0.1077501870 -0.4381438789 0.0005002154

siblings computertime exercisehours musiccds playgames

0.7107673587 0.1901958654 0.0460659418 0.0033897185 0.1434152000

Sigma^2 = 24.18417

The residulas seem centered in zero and homoschedastic and we can assume the hypothesis of normality (we perform a Shapiro test and the pvalue is 0.01278 so at level 1% we accept)

1. Estimated coeff:

Intercept 4.0894018365

age 0.0494139480

height .

dummy\_sex .

distance 0.0003598352

siblings 0.4526208912

computertime 0.1422680784

exercisehours .

musiccds 0.0012825007

playgames 0.0230406929

where the dot indicates that the associated coeff is practically = 0 so it is not significant, moreover we can see that also the coeff of distance and musiccds are near to 0 so we could perform a test to check if we can remove them -> Non chiede di interpretare!

1. Setting lambda via cross-validation we obtain that the best value of lambda, ie the one which minimizes the CV error, which is 0.4328761.

* ATT: Non devo scegliere per forza il minimo, guarda la riduzione delle variabili! Muoviti nel range della standard deviation se ti permette di scartare altre var
* Qui scelgo lambda = 1

Da aggiornare:

Choosing this value the parameters are:

Intercept = 5.7498287211

Siblings = 0.3365513849

Computertime = 0.1231026862

Musiccds = 0.0003359460

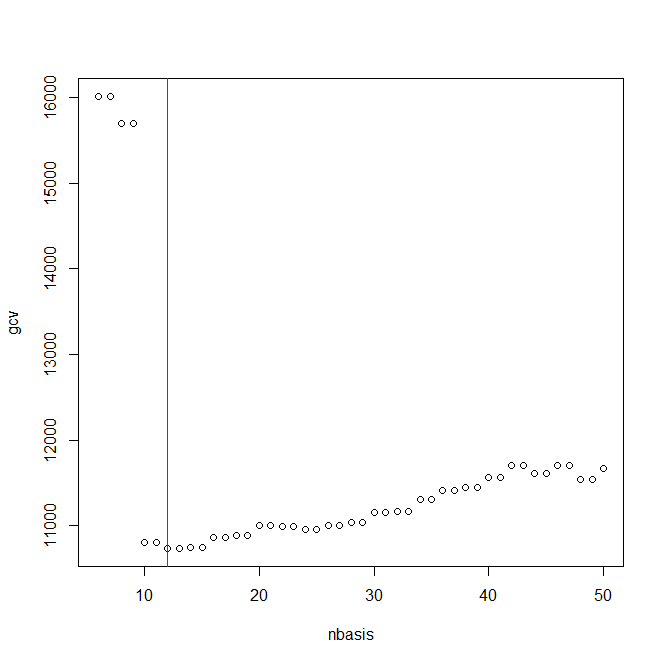
The other coeff associated with the other variables are null.

1. The pointwise estimate is = 7.358221 questo usando I coeff di lasso

Q: uso lasso o devo rifittare un modello di regressione lineare con solo le variabili selezionate da lasso?

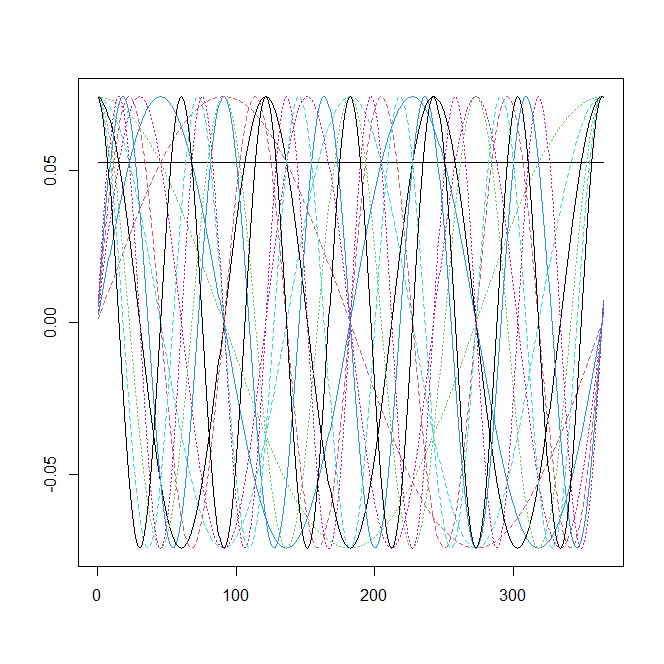
Ex4)

Perform GCV to choose the optimal number of basis function

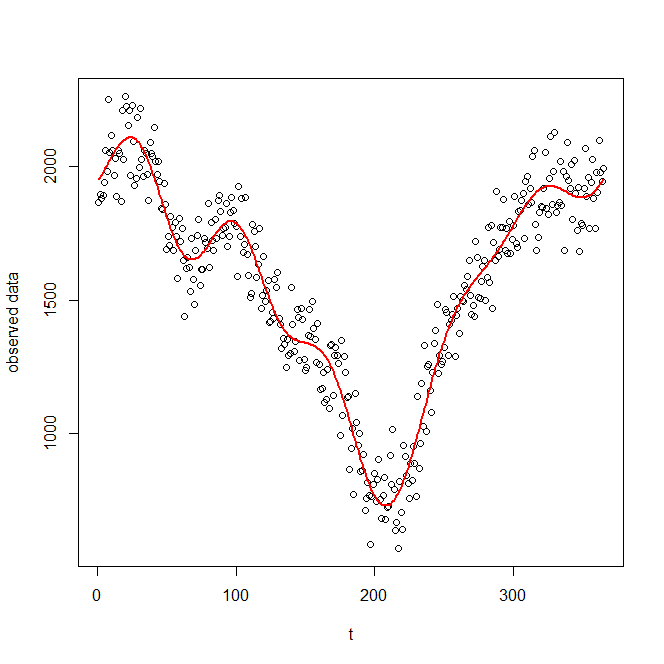


We choose nbasis = 12

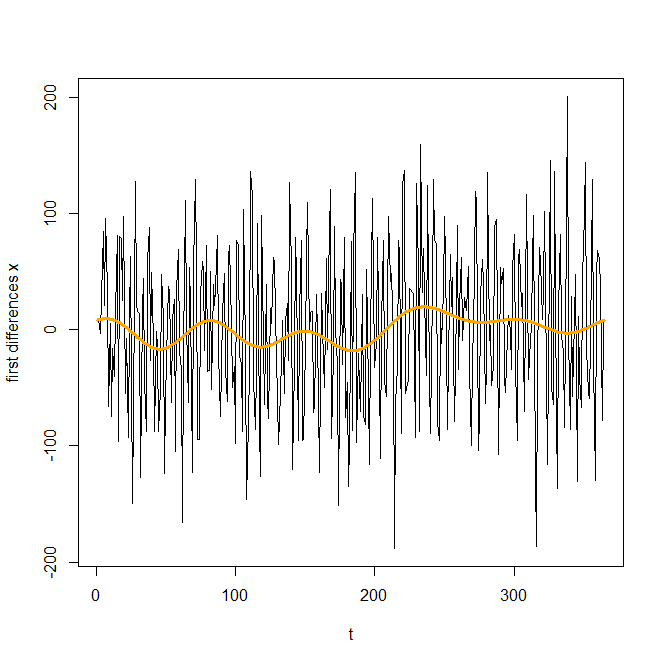
Basis :



Smooth data:

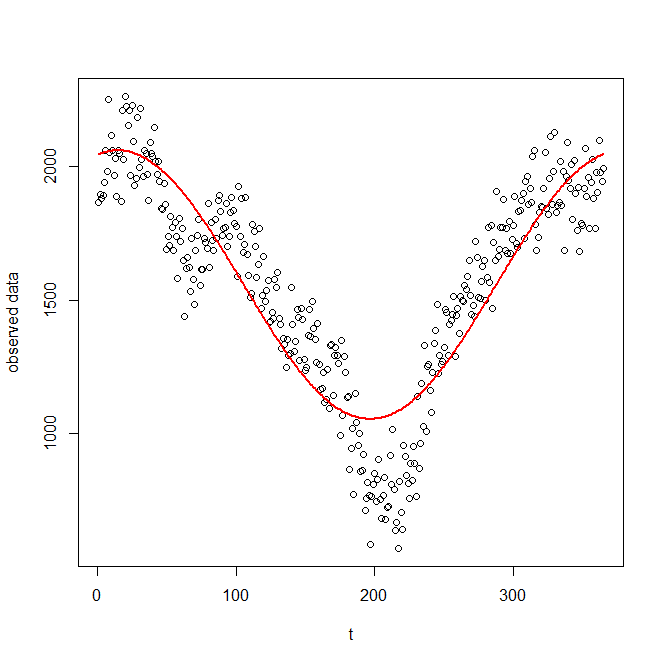


1. First derivative:



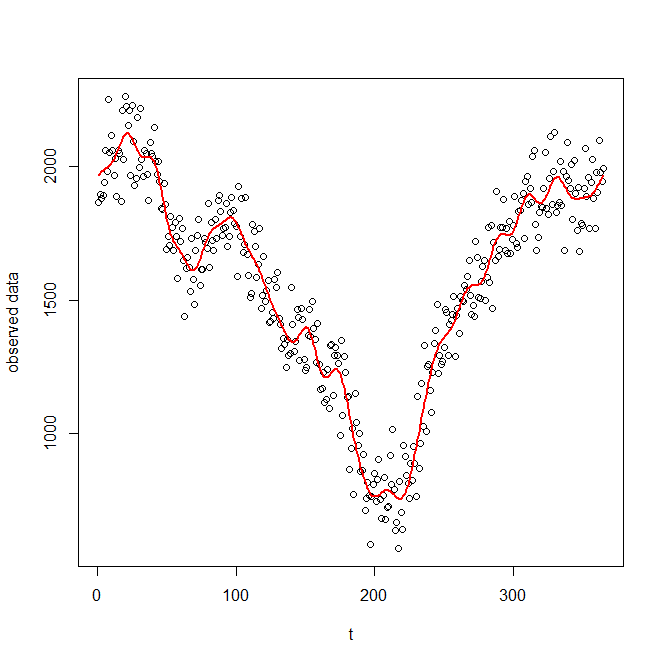
We can observe that the first derivative estimated from the data is really noisy but the smoothed one is still good.

1. The phenomenon of oversmoothing = underfitting appears when the chosen number of basis is too low. To see well the effect of the smoothed data we can choose nbasis = 3 and obtain



We see that the curve is not capturing the real behaviour of the data and it is smoothing too much peaks and cavities.

1. The phenomenon of undersmoothing= overfitting appears when the chosen number of basis is too large. To see well the effect of the smoothed data we can choose nbasis = 40 and obtain



Here we see that the smoothed function is trying to interpolate the data and so also the noise and this is not the behvaiour that we want. Indeed we can also see from the plot of the GCV that with nbasis = 40 the error is increased a lot wrt the optimal case.