TDE 03/07/17

Problema 1.

a) Il modello ANOVA che andiamo a ipotizzare è un modello ANOVA con due fattori. Il primo è legato alla città e il secondo al tipo, il modello che creiamo e di conseguenza:
   Xijk= MU + TAU i + Beta j + gamma ij + epsilon ijk

   Dove Mu indica la media totale Tau l'effetto di essere in città diverse Beta l'effetto dovuto ai tipi diversi e infine gamma rappresenta l'interazione tra i due fattori esposti sopra. Per assunzione epsilon devono essere normali indipendenti e identicamente distribuiti con media zero e varianza comune sigma.
   Le assunzioni sono verificate attraverso uno shapiro test nei singoli gruppi per la normalità e un bartlet test per verificare l'uguaglianza delle varianze

b) Come test per valutare la possibilità di rimuovere alcune covariante utilizziamo prima il test sulle interazioni quindi abbiamo come ipotesi nulla che gamma ij =0 e come H1: gammij!=0 E costruiamo il test basandoci sulla statistica test SS interaction/(g-1)*(b-1)/ SSres/(gb)n-1)) se questa è maggiore di F1-alpha ((g-1)*(b-1),gb(n-i) allra l'ipotesi nulla è rifiutata. Nel nostro case sia b che g sono pari a 2. Da summary dell'ANOVA evinciamo che le interazione possono essere trascurate e procediamo con un modello additivo a questo punto ci basiamo su una statistica test SStret1/(g-1)/SSres+SSinteracton/gbn-b-g+1 allo stesso modo rifiutiamo se questa è maggiore di F1-alpha ((g-1),gbn-b-g+1). Dal summary dell'ANOVA possiamo considerare anche l'effetto del trattamento relativo alla città nullo quindi il nostro modello ridotto sarà del tipo
   Xijk= MU + Beta j + epsilon ijk dove le epsilon devono essere normali indipendenti e identicamente distribuiti con media zero e varianza comune sigma.
c) Dobbiamo ora valutare gli intervalli di Bonferroni per la differenza delle medie in base al tipo di abito. In particolare dobbiamo fornire un solo intervallo di confidenza
   IC 0.95 (beta1 – beta2) = [media campionaria 1- media campionaria2 +- (t0.95 (n-2)) * sqrt(SSres^2/ n-2)(1/n1 +1/n2)= [14.66669 15.15225]

Problema 2.

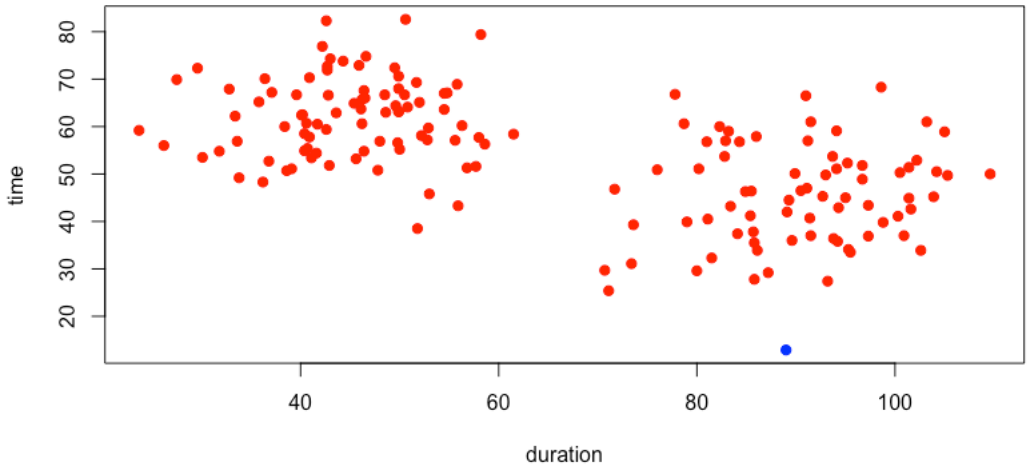
a) Risolviamo il problema nel contesto dei Paired Data. In particolare abbiamo che le due misurazioni sono fatte sulle stessa unità statistica e quindi dobbiamo andare a valutare le differenze nelle due misurazioni. Definiti Di = Xi1-Xi2 ipotizziamo che le differenze siano normali e lo verifichiamo con un mcshapiro test. A questo punto possiamo valutare la differenza tra le due media nelle diverse variabili basandoci su un test per la media di una normale multivariata
   H0: mu =0 vs H1: mu!=0
   Lo basiamo sulla statistica test m*(Dsample)*S-1*(Dsample) e valutiamo il p-value del nostro test. Il p-value dalla computazione è nullo di conseguenza rifiutiamo l'ipotesi nulla.
b) Per avere i quattro intervalli di confidenza ci rifacciamo alla stessa quantità pivotale di prima in particolare avremo che
   SimCI0.95(mui) = [(Dsample)i +- sqrt((n-1)p/n-p F0.95(p,n-p) *(Sd)ii/n)]

	inf	center	sup
rice	-37.34390	-2.145937	33.05202]
sashimi	107.51324	123.577812	139.64238]
vegetables	-12.40493	11.463438	35.33181]
okashi	111.98999	118.150000	124.31001]

Possiamo vedere che sia l'intervallo di confidenza del riso che delle verdure contiene lo zero di conseguenza non abbiamo evidenza statistica per dire che il loro consumo sia diverso in diversi momenti. Mentre il consumo di sashimi e di okashi aumenta di molto nel periodo di festa.

Problema 3

a) Vogliamo creare un clustering riguardante il successo o meno di un tour in abse all covariate e caraterristice del tour che conosciamo. In particolare selezioniamo come distanza tra due unità statostistiche la loro distanza euclidiana e coe distanza tra i cluster sfruttiamo il metodo single linkage una volta definite queste due quantità possimao direttamente applicare un algoritmo di clustering ierarchico agglomerativo.



Where the centers are:

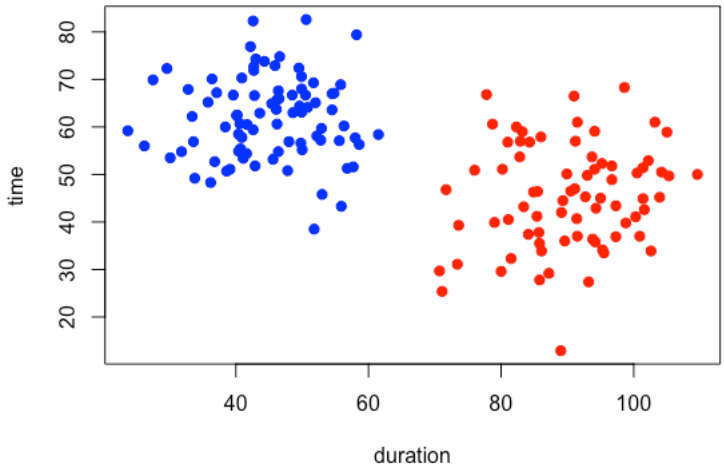
- For the red cluster: (66.07925, 54.33019)
- For the blue cluster (89, 12.9)

The numerosity of the clusters are:

- For the red cluster: 159
- For the blue cluster: 1

The cophenetic coefficient is:
0.8781562

b) The clustering prodice at point a is unsatofactory since the sigle linkage is not good when the two cluster are not well separated we can help and have a better cluster with avagare or complete linkage. In this case :



Where the centers are:

- For the red cluster: (90.23867, 45.20000)
- For the blue cluster (45.03176, 61.89882)

The numerosity of the clusters are:

- For the red cluster: 75
- For the blue cluster: 85

The cophenetic coefficient is:
0.8807631

c) Una volta divisi I dati nei due gruppi possiamo andare a controllare se sono normali e se hanno la stessa covarianza. In particolare grazie a un mc.shapiro test possiamo vedere che entrambi i gruppi sono normali ma non hanno la stessa covarianze quindi supponiamo che n sia abbastanza grande per valutare questi intervalli in modo asintotico
SimCI 0.95(mu1-mu2)i = [sample mean 1- sample mean 2 +- z 1-0.05/4 * sqrt(1/n1 *sigma 1 +1N2*sigma 2)ii]
BonfCI 0.95 (mu1) = [sample mean 1+- t 1-0.05/4 (n1-1) * sqrt(1/n1 *sigma 1)ii]

Diff.mean.duration 42.12192 48.29188
Diff.mean.starting time -20.21498 -13.18266
Mean.duration 87.83011 92.64723
Mean.Starting time 42.31408 48.08592

d) Quello che possiamo notare dagli intervalli di confidenza è che tendenzialmente tour che durano di più e iniziamo prima hanno maggior successo. Di conseguenza una strategia di successo potrebbe essere quella di valutare tour che partano alle 16:45 e durino 90 minuti.

Problema 4

a) We based on the model
$E = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \varepsilon$,
with E the extension of the garden, x_1,x_2,x_3,x_4 the number of carps, maple trees, cherry trees and stones respectively, and $\varepsilon \sim N(0,\sigma^2)$.

In particular our design matrix is formed by Z= [1 x_1 x_2 x_3 $x_{4.}$] the estimation of beta is performed by OLS so that

Beta = (Z'Z)^(-1)Z' Y

Were Y is the vector of the observed response so E.

(Intercept) carps maple cherry stones

1442.86188 16.34107 28.69118 13.12381 14.02307

The residuals are normal and has a good cloud shape

b) We need to perform a test of significance on linear combination of the beta and in particoluar two of them at the time so we perform to test like:

H0: C*beta=0 vs H1: C*beta !=0

We use as test statistic (C*beta)' *(C(Z'Z)^(-1)*(C*beta)/p*S^2 and we compute the pvalue of the test.

If C =(0 0 1 0 0

0 0 0 1 0)

Then the p-value is really small so that we reject H0. The same hold for

C = [0 1 0 0 0

0 0 0 0 1]

c) The wekness of this model is collinearity indees if we compute the vif of all the regressors we get

carps maple cherry stones

19.92811 18.87960 15.35421 17.45596

They are all grater then 10 so we have a lot of collinearity in the data. One approach to avoid this problematic is to use PCA regression

So we first find the PC that are orthogonal and then we base our analysis on the scores of the data. Computing the PCA we can reduce the dimensionality to the first two PC. The first refer to a general weighed sum while the second to contrast between lake element and ground element.

Given b0,b1,b2 form the PCA regression we need to recompute the value of the beta for the initial model

y= b0 + b1* PC1 + b2* PC2 + eps = b0 + b1*(e11*(X1-m1)+e21*(X2-m2)+e31*(X3-m3)+ e41*(X4-m4)) + b2*(e12*(X1-m1)+e22*(X2-m2)+ e32*(X3-m3)+ e42*(X4-m4)) + eps =

= b0 - b1*e11*m1 - b2*e12*m1 - b1*e21*m2 - b2*e22*m2 - b1*e31*m3 - b2*e32*m3 - b1*e41*m4 – b2*e42*m4 + (b1*e11+b2*e12)*X1 + (b1*e21+b2*e22)*X2 +(b1*e31+b2*e32)*X3+(b1*e41+b2*e42)*X4+ eps

(Intercept)

1484.613

Carps

15.64089

Maple

19.25401

Cherry

21.3864

Stones

15.08058