Prima Relazione di Statistica

Regressione Lineare e Analisi delle Componenti Principali

Manni Matteo

PRESENTAZIONE DEL PROBLEMA

Immaginiamo che il problema prefissato sia quello di cercare di capire quali sono i fattori che determinano i costi di gestione di una struttura bibliotecaria e provare a ricavare un modello che preveda e quantifichi i costi sulla base di alcuni parametri.

La relazione si svilupperà sullo studio di dati raccolti in alcune regioni italiane riguardanti il sistema bibliotecario pubblico italiano.

In particolare, per capire quali sono i fattori rilevanti per il problema seguiremo l'analisi delle componenti principali, mentre per quanto riguarda la previsione dei costi di gestione proveremo ad analizzare l'adeguatezza di modelli di regressione lineare.

DATASET

Per le analisi svolte è stato utilizzato un dataset reperito dal sito dell'Istat, che ha come campioni i dati raccolti in diverse regioni italiane. Link:
http://dati.istat.it/viewhtml.aspx?il=blank&vh=0000&vf=0&vcq=1100&graph=0&view-metadata=1&lang=it&QueryId=22037&metadata=DCIS_BIBLIOT#

È stata messa insieme la tabella del 2009 con quella del 2010. Il dataset è composto in totale da 9 fattori e 28 osservazioni, i fattori diventano però 7 se escludiamo quelli categorici che indicano la regione e l'anno.

1. **territorio**: la regione italiana dalla quale provengono i dati
2. **anno**: l'anno nel quale sono stati raccolti i dati
3. **lettori_ita**: il numero di lettori italiani
4. **lettori_str**: il numero di lettori stranieri
5. **opere_consultate**: il numero di opere consultate
6. **lettori**: il numero di lettori totali (**lettori_ita** + **lettori_str**)
7. **posti_per_lettori**: il numero di posti per i lettori
8. **persone_ammesse_al_prestito**: il numero di persone ammesse al prestito bibliotecario
9. **spese_gestione**: i costi di gestione in euro

Il numero delle osservazioni (28) non è particolarmente grande, questo fatto potrebbe compromettere alcune delle analisi che andremo a sviluppare. Sarebbe più opportuno avere una maggiore quantità di dati.

ANALISI

ANALISI DELLE CORRELAZIONI

Dopo aver caricato in memoria la tabella escludiamo dall'analisi le prime due colonne (**territorio** e **anno**) perché non utili agli scopi. Possiamo adesso dare una prima valutazione alle correlazioni tra i fattori.

GRAFICI DI CORRELAZIONE

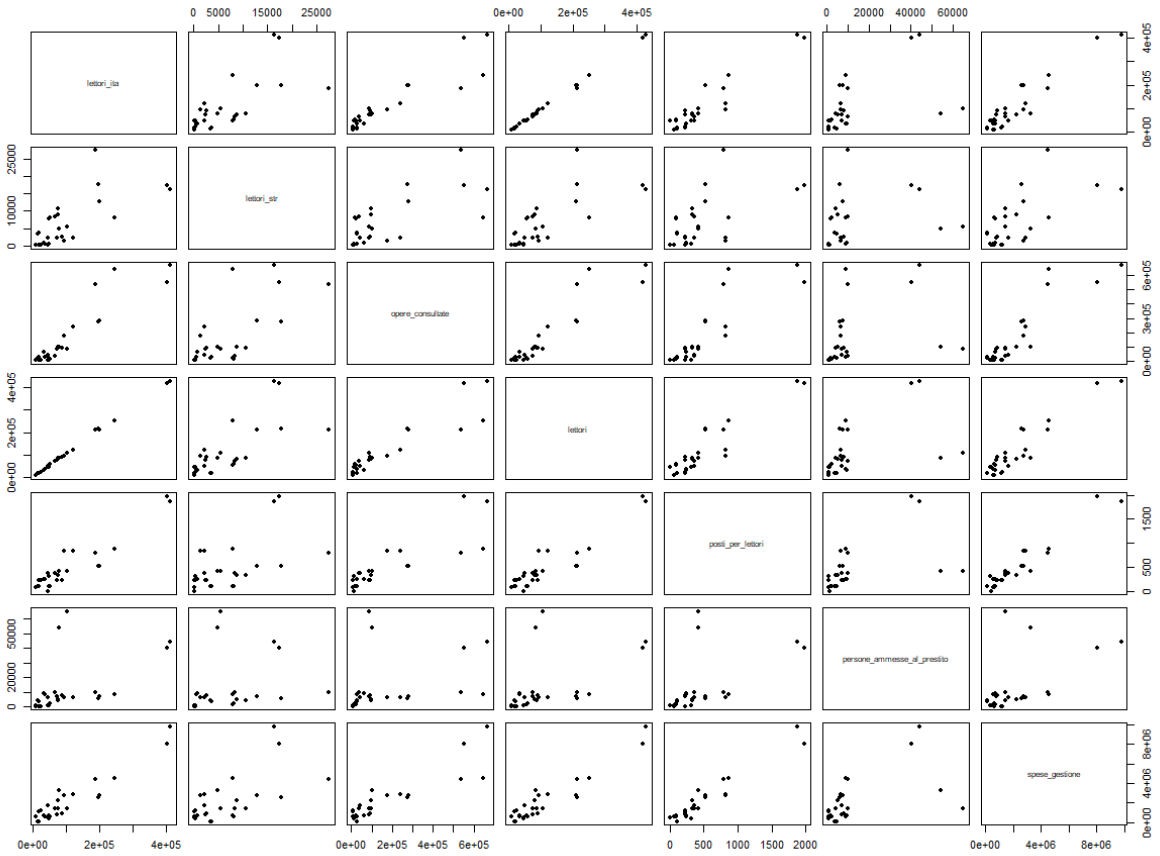

fig.1

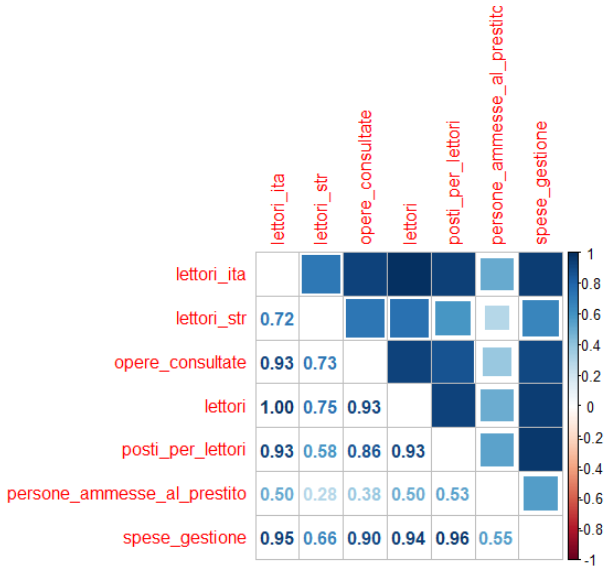

fig.2

Come potevamo aspettarci vediamo dalla fig.1 che tra i fattori **lettori** e **lettori_ita** c'è una forte correlazione. Potevamo immaginare lo stesso comportamento (o simile) tra i fattori **lettori** e **lettori_str**; in realtà tra questi ultimi la correlazione, se pur presente, non è estremamente marcata come nel primo caso.

Dalla fig.2 ci accorgiamo che stiamo trattando una tabella ben correlata. I fattori hanno quasi tutti correlazioni forti tra loro. L'unico fattore che si correla meno agli altri è **persone_ammesse_al_prestito**.

Da questa prima acerba analisi ipotizziamo che i costi di gestione non siano fortemente influenzati dal numero di persone ammesse al prestito (corr<60).

ANALISI DELLE COMPONENTI PRINCIPALI

Dall'analisi delle componenti principali cerchiamo di ricavare informazioni utili al problema interpretando la composizione delle componenti.

Per avere un risultato significativo sarà necessario rimuovere il fattore **spese_gestione** e normalizzare la tabella.

STUDIO DELLA VARIANZA

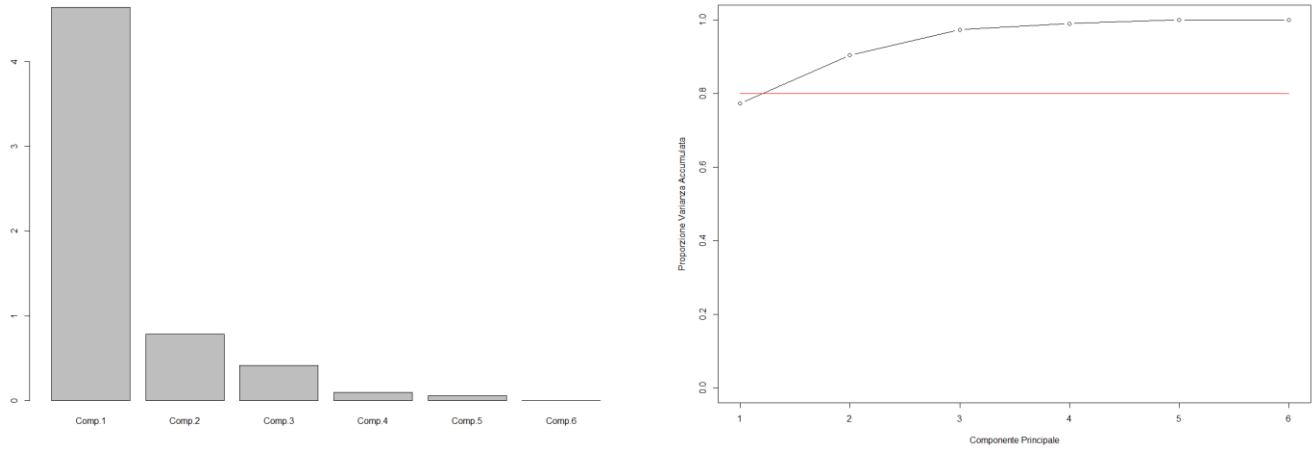

fig.3

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5       Comp.6
Standard deviation     2.1532808 0.8883941 0.6437900 0.31497181 0.24589600 9.991226e-09
Proportion of Variance 0.7727697 0.1315407 0.0690776 0.01653454 0.01007747 1.663743e-17
Cumulative Proportion  0.7727697 0.9043104 0.9733880 0.98992253 1.00000000 1.000000e+00
```

Dallo studio della Varianza si nota che anche solo con la prima componente si raggiunge un ottimo 77%. Potrebbe quasi bastare solo quella per sforare la soglia empirica dell'80% che però si raggiunge e supera con la seconda componente.

Sono sufficienti quindi le prime due componenti per catturare il problema il che è ottimo anche per l'interpretabilità.

COMPOSIZIONE DELLE COMPONENTI

Se analizziamo la matrice dei loadings (fig.4) concentrandoci sulle prime due componenti possiamo chiaramente assegnare alla prima le caratteristiche **lettori_ita**, **opere_consultate**, **lettori** e **posti_per_lettori**. La seconda componente invece è caratterizzata principalmente da **persone_ammesse_al_prestito**. Rimane incerto il fattore **lettori_str** che però dopo un'opportuna rotazione (fig.5) possiamo attribuire alla prima componente.

La prima componente è quindi caratterizzata da fattori che riguardano la mole di lettori della biblioteca e potremmo forse considerarla una componente riassuntiva, andiamo però avanti con le considerazioni.

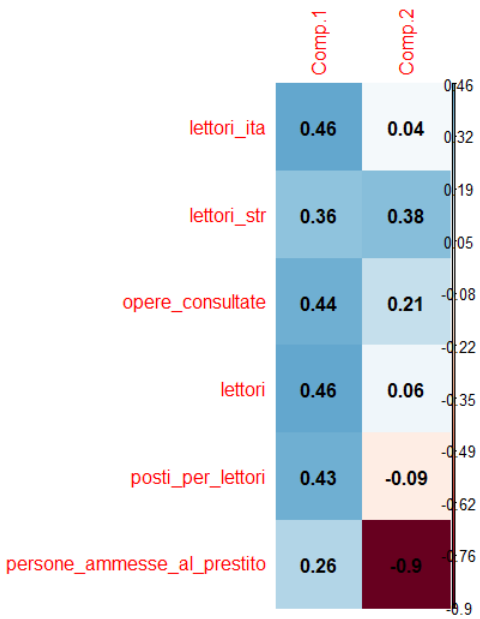

fig.4

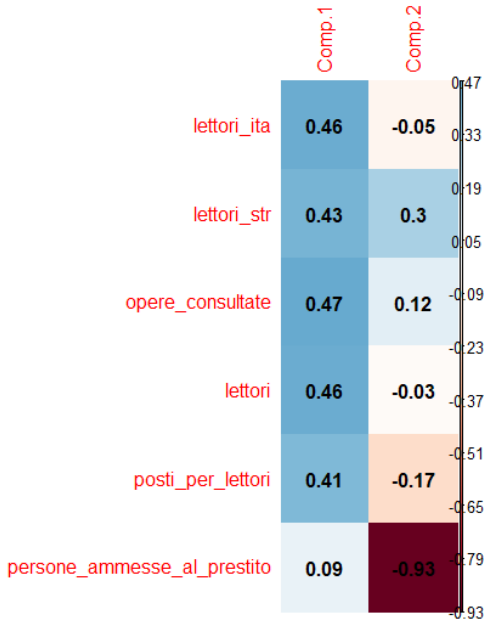

fig.5

BIPLOT

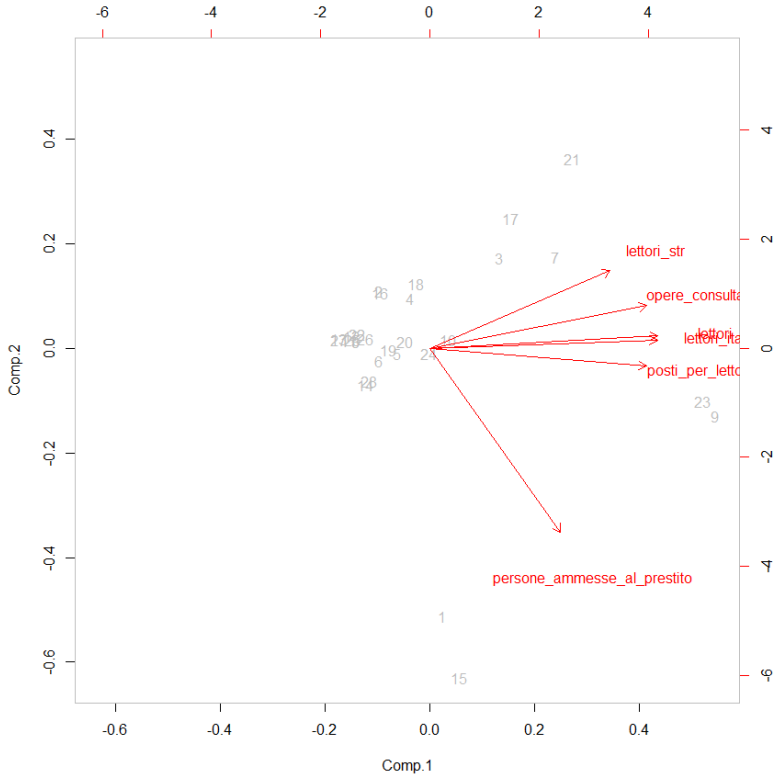

fig.6

Sul piano principale possiamo vedere ciò che abbiamo in parte già discusso. Ad eccezione di **persone_ammesse_al_prestito**, notiamo che più o meno tutte le frecce dei fattori sono allineate alla prima componente principale.

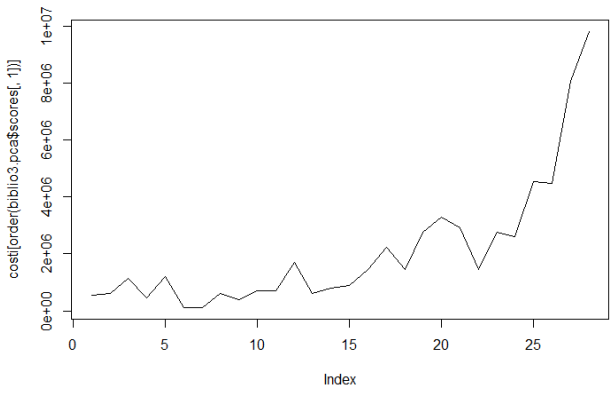

fig.7

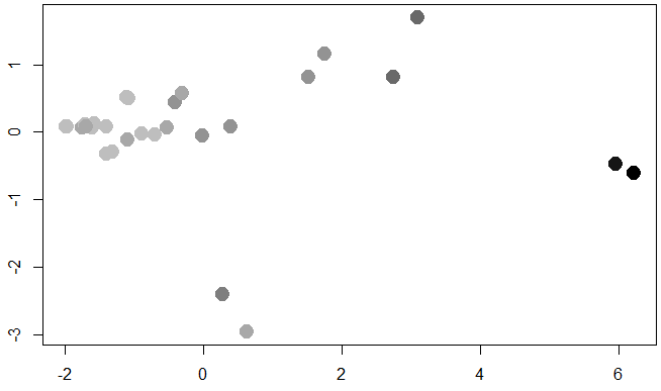

fig.8

Possiamo creare un grafico (fig.7) avente sull'asse delle ascisse i campioni ordinati in base agli 'scores' (della prima componente principale) dei campioni stessi, mentre sull'asse delle ordinate i costi di gestione rispettivi di ogni campione. Si evincere che effettivamente i costi crescono al crescere della prima componente principale, il che significa che possiamo farci un'idea delle spese di gestione in base al posizionamento sul piano.

Per convincerci ancora di più di questo fatto possiamo visualizzare come si collocano i campioni sul piano principale, colorando con una scala di grigi i campioni in base al loro rispettivo fattore di **spese_gestione**.

Se osserviamo i punti e la loro tendenza nell'essere più scuri quando si trovano nella parte destra del grafico, concludiamo esattamente quello che abbiamo già discusso.

CONCLUSIONI

Con la prima componente principale che quasi si allinea perfettamente al fattore **lettori** possiamo avere un'idea di quello che è **spese_gestione**, essendo anch'essa allineata. Possiamo considerare inoltre la prima componente come riassuntiva per il nostro problema.

REGRESSIONE

Come già anticipato nell'introduzione abbiamo lo scopo di riuscire a prevedere le spese di gestione sulla base degli altri fattori. A tal fine analizzeremo i risultati proposti da modelli di regressione.

Dall'analisi delle correlazioni avevamo visto che **spese_gestione** ha correlazioni forti con **lettori** e **posti_per_lettori**. Possiamo innanzitutto provare a costruire due modelli di regressione lineare semplice con questi due fattori.

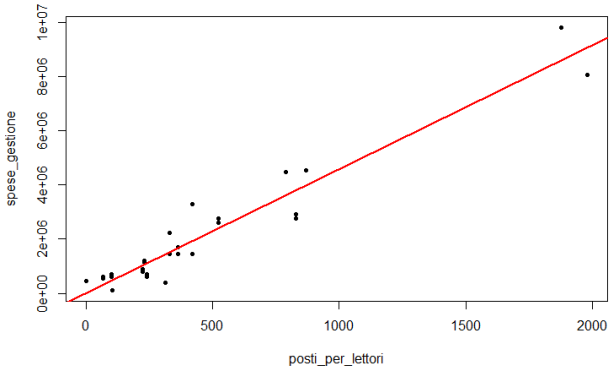

fig.9

```
Adjusted R-squared: 0.9231

p-value: 3.23e-16
```

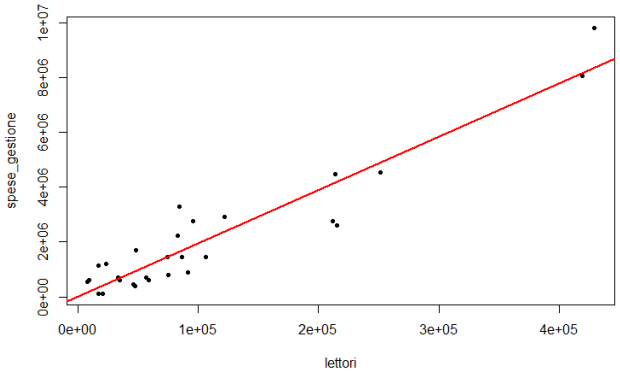

fig.10

```
Adjusted R-squared: 0.8853

p-value: 5.983e-14
```

Il risultato non è affatto pessimo. Nel caso del modello con **posti_per_lettori** si raggiunge una varianza spiegata aggiustata superiore al 92%, il che è ottimo. Per quanto riguarda il modello lettori si arriva ad un buon 88%. In entrambi i casi il p-value è tendente allo 0.

REGRESSIONE LINEARE MULTIVARIATA

Vediamo adesso un modello di regressione multivariata, anche se potremmo quasi valutare di fermarci dato gli ottimi risultati ottenuti dalla Regressione Semplice. Inoltre, è ragionevole pensare che il modello multivariato dovrà essere ridotto.

```
Residuals:
    Min      1Q  Median      3Q     Max
-995171 -313678  -95612  369667 1133369

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -48558.314 157949.388  -0.307 0.761408
lettori_ita                  1.835      4.064   0.451 0.656053
lettori_str                 17.850     24.302   0.735 0.470396
opere_consultate             2.446      1.469   1.665 0.110007
lettori                         NA         NA      NA       NA
posti_per_lettori         2938.610    667.064   4.405 0.000224 ***
persone_ammesse_al_prestito   11.883      7.604   1.563 0.132390
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 546600 on 22 degrees of freedom
Multiple R-squared:  0.9539,    Adjusted R-squared:  0.9435
F-statistic: 91.15 on 5 and 22 DF,  p-value: 5.969e-14
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-786009 -378929   -8131  257435 1478629

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.420e+04  1.469e+05   0.369  0.71518
posti_per_lettori 3.447e+03  4.435e+02   7.774 3.95e-08 ***
opere_consultate  3.103e+00  1.056e+00   2.939  0.00699 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 560500 on 25 degrees of freedom
Multiple R-squared:  0.945,     Adjusted R-squared:  0.9406
F-statistic: 214.6 on 2 and 25 DF,  p-value: < 2.2e-16
```

Otteniamo una varianza spiegata aggiustata del 94% e un p-value tendente allo 0. Abbiamo però un p-value dei fattori molto elevato. Quindi proviamo a ridurre il modello togliendo uno ad uno i fattori per grandezza di p-value, arriviamo così ad un modello ridotto con solo due fattori (**posti_per_lettori** e **opere consultate**), la cui varianza spiegata aggiustata però non scende sotto il 94%.

Vediamo che il fattore più significativo è **posti_per_lettori**, lo stesso sul quale abbiamo creato il modello di regressione semplice in precedenza. Rispetto a quest'ultimo, non si ha un incremento della varianza spiegata aggiustata così sostanziale che giustifichi l'incremento di complessità.

PREDIZIONI PRODOTTE DAI MODELLI A CONFRONTO

Proviamo a capire se l'aumento di complessità può essere giustificato analizzando le predizioni prodotte dai due modelli. Per 500 iterazioni dividiamo il data_set in training_set e test_set e ricaviamo un grafico degli errori medi riguardo le predizioni prodotte dai due modelli. In particolare, due linee: una blu per il modello di regressione semplice e una rossa per il modello di regressione multivariata.

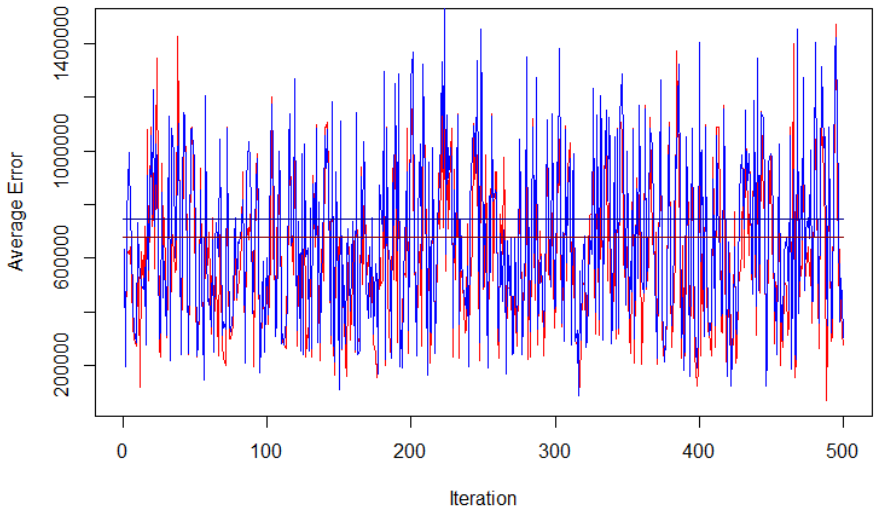

fig.11

Dopo una prima 'occhiata' le due linee sembrano oscillare nello stesso range di valori, il che potrebbe far pensare che tra i due modelli non ci sia una sostanziale differenza. In realtà se esaminiamo la media dei valori

delle due linee vediamo che gli errori del modello di regressione semplice sono più grandi (come potevamo banalmente prevedere). Se quantifichiamo la differenza vediamo che il modello di regressione multivariata è più accurato di circa il 10%. Non c'è una differenza netta ma comunque abbastanza apprezzabile da farci scegliere il modello di regressione multivariata con i due fattori **posti_per_lettori** e **opere_consultate**.

ANALISI DEI RESIDUI

Da un primo plot dei residui (fig.12) vediamo che questo non ci fornisce informazioni sostanziali, il motivo è che probabilmente ci sono troppi pochi campioni. In ogni caso sembra che i residui siano abbastanza distribuiti attorno allo 0 e non presentino stranezze evidenti. Se osserviamo la densità discreta ci accorgiamo che le cose non sembrano andare come vorremmo, la causa anche qui potrebbe riguardare i pochi campioni raccolti. Per capire meglio possiamo raffigurare nello stesso grafico (fig.13) la densità empirica (in rosso) e la densità gaussiana (in nero). Vediamo che le due linee non si discostano in realtà in maniera così drammatica. Ci convinciamo di questa ultima considerazione analizzando quantitativamente le differenze tramite un grafico Quantile-Quantile (fig.14); dal quale vediamo che i punti hanno una buona aderenza alla retta rossa tra -1.5 e 1.5.

Possiamo considerare questi risultati accettabili.

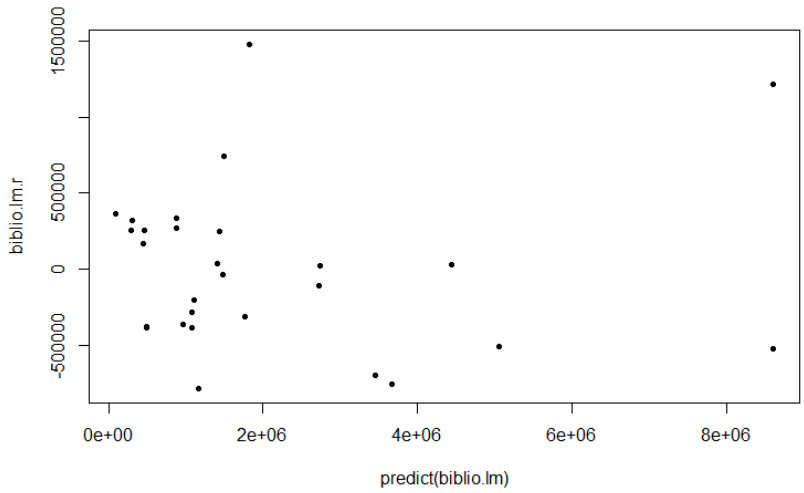

fig.12

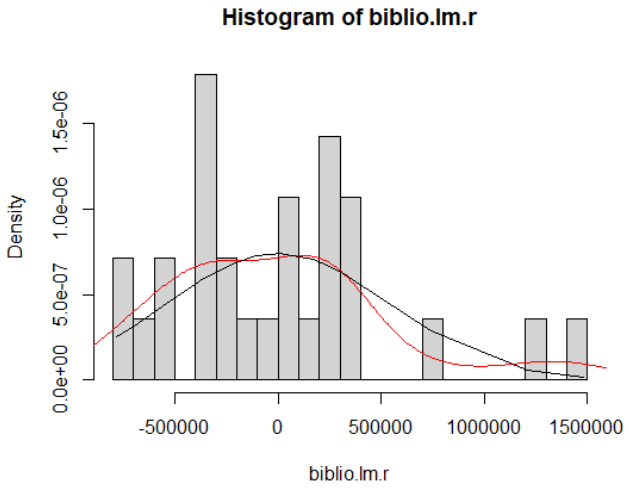

fig.13

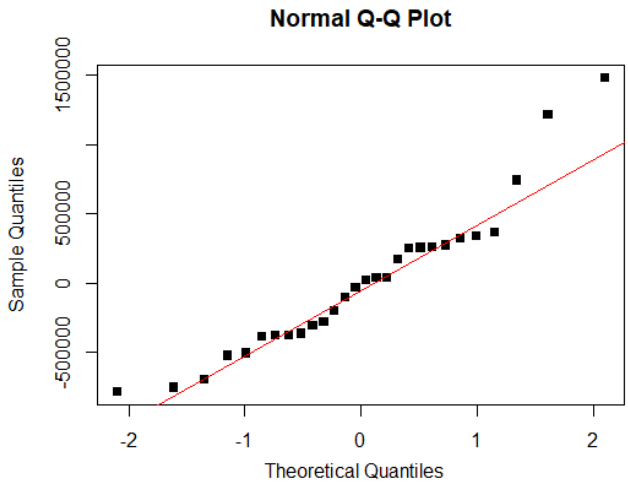

fig.14

CONCLUSIONI CON STIMA DELLE INCERTEZZE NELLA PREVISIONE

Concludiamo l'analisi con una stima delle incertezze delle previsioni, senza la quale non riusciremmo a dare dei risultati significativi. Vediamo nella fig.15 che le previsioni del modello (in giallo) rientrano all'interno dei margini di predizione (in blu); rientrano anche all'interno dei margini di confidenza empirici (in verde) e dei margini di confidenza parametrici (in rosso).

Quindi possiamo dire che il modello, avente come ingresso il numero dei posti per i lettori e il numero di opere consultate, risulta adatto per la previsione dei costi di gestione.

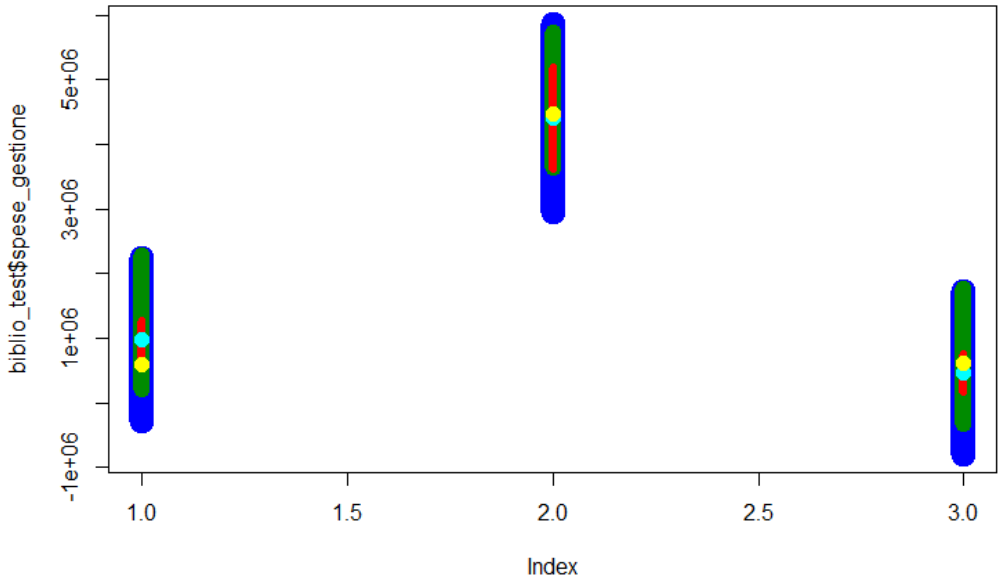

Fig.15

APPENDICE

CODICE R UTILIZZATO PER LE ANALISI

```
#caricamento in memoria
biblio <- read.csv("tabella.csv")
biblio2 = biblio[,-1:-2]

#correlazione
library(corrplot)
plot(biblio2,pch=20)
corrplot(cor(biblio2),"square")
corrplot.mixed(cor(biblio2),lower = "number",upper = "square", tl.pos = "lt")

#PCA
biblio3 = biblio2[,-7]  #tolgo spese_gestione
biblio3.pca = princomp(biblio3, cor=T)
summary(biblio3.pca)
plot(biblio3.pca, main = "")
plot(cumsum(biblio3.pca$sdev ^ 2) / sum(biblio3.pca$sdev ^ 2),type = "b",ylim = c(0, 1),xlab =
"Principal Components",ylab = "Percentage Cumulative Variance")
segments(1, 0.9, 6, col = "red")

#loading senza rotazione
biblio.ld=loadings(biblio3.pca)
corrplot(biblio.ld[,1:2], is.corr=FALSE, method="color",addCoef.col = "black", number.digits = 2)

#loadings con rotazione
biblio.rot=varimax(biblio.ld[,1:2])$loadings
```

```
corrplot(biblio.rot, is.corr=FALSE, method="color",addCoef.col = "black", number.digits = 2)
#biplot
biplot(biblio3.pca, col=c("gray","red"))

#componente riassuntiva
costi=biblio[,9]
plot(costi[order(biblio3.pca$scores[,1])],type="l")

#plot delle componenti dei record sul piano principale
grad=colorRampPalette(c("grey","black"))
scol=grad(10)
x=(costi-min(costi))/(max(costi)-min(costi))
sidx=1+floor(10*0.99*x)
pred=biblio3.pca$scores[,1:2]
plot(pred,pch=19,cex=2,col=scol[sidx],main=ncol)

#REGRESSIONE
#regressione lineare con lettori
biblio.lm=lm(spese_gestione~lettori,data=biblio2)
plot(biblio2$lettori,biblio2$spese_gestione,pch=20)
abline(biblio.lm,lwd=2,col="red")
summary(biblio.lm)

#regressione lineare con posti_per_lettori
biblio.lm=lm(spese_gestione~posti_per_lettori,data=biblio2)
plot(biblio2$posti_per_lettori,biblio2$spese_gestione,pch=20,xlab = "posti_per_lettori",ylab =
"spese_gestione")
abline(biblio.lm,lwd=2,col="red")
summary(biblio.lm)

#regressione multivariata
biblio.lm=lm(spese_gestione ~ .,data=biblio2)
summary(biblio.lm)

#riduzione del modello
biblio.lm=lm(spese_gestione ~ posti_per_lettori + opere_consultate,data=biblio2)
summary(biblio.lm)

#previsione
m=500
semp_err=rep(0,m)
multi_err=rep(0,m)

for(i in 1:m){
  idx=sample(28,3)
  biblio_train=biblio2[-idx,]
  biblio_test=biblio2[idx,]
  biblio_train.lm.semp=lm(spese_gestione ~ posti_per_lettori, data=biblio_train)
  biblio_train.lm.multi=lm(spese_gestione ~ posti_per_lettori + opere_consultate,
data=biblio_train)

  biblio.lm.p.semp=predict(biblio_train.lm.semp,biblio_test)
  biblio.lm.p.multi=predict(biblio_train.lm.multi,biblio_test)

  semp_err[i] = mean((biblio.lm.p.semp - biblio_test$spese_gestione)^2)
  multi_err[i] = mean((biblio.lm.p.multi - biblio_test$spese_gestione)^2)
}

sqrt(mean(semp_err))
sqrt(mean(multi_err))

plot(sqrt(multi_err),type = "l", col = "red", xlab = "Iteration", ylab = "Average Error")
```

```
lines(sqrt(semp_err), col = "blue")
segments(0, sqrt(mean(multi_err)), m, sqrt(mean(multi_err)), col = "darkRed")
segments(0, sqrt(mean(semp_err)), m, sqrt(mean(semp_err)), col = "darkBlue")

(100/sqrt(mean(multi_err)))*sqrt(mean(semp_err)) #percentuale di differenza

#residui
biblio.lm=lm(spese_gestione ~ posti_per_lettori + opere_consultate,data=biblio2)
biblio.lm.r=residuals(biblio.lm)
plot(predict(biblio.lm),biblio.lm.r,pch=20)

plot(fitted(biblio.lm),biblio.lm.r,pch=20)
hist(biblio.lm.r,20,freq=F)

hist(biblio.lm.r,freq=F,20)
lines(density(biblio.lm.r),col="red")
lines(sort(biblio.lm.r),dnorm(sort(biblio.lm.r),m=mean(biblio.lm.r),sd(biblio.lm.r)))

qqnorm(biblio.lm.r, pch=15)
qqline(biblio.lm.r, col="red")

#Stima delle incertezze nella previsione
alpha=0.95
idx=sample(28,3)
biblio_train=biblio2[-idx,]
biblio_test=biblio2[idx,]
biblio_train.lm=lm(spese_gestione ~ posti_per_lettori + opere_consultate,data=biblio_train)

#intervalli di confidenza
biblio_test.ci=predict(biblio_train.lm,biblio_test,interval="confidence",level=alpha)
#intervalli di predizione
biblio_test.pi=predict(biblio_train.lm,biblio_test,interval="prediction",level=alpha)
#intervalli di confidenza empirici
biblio_train.r=resid(biblio_train.lm)
qi=quantile(biblio_train.r,(1-alpha)/2)
qs=quantile(biblio_train.r,(1+alpha)/2)


ymin=min(c(biblio_test.pi[,2],biblio_test.ci[,2]))
ymax=max(c(biblio_test.pi[,3],biblio_test.ci[,3]))
plot(biblio_test$spese_gestione,ylim=c(ymin,ymax))
x=1:3

#intervalli di predizione
segments(x,biblio_test.pi[,2],x,biblio_test.pi[,3],col="blue",lwd=18)
#intervalli di confidenza empirici
segments(x,biblio_test.pi[,1]+qi,x,biblio_test.pi[,1]+qs,col="green4",lwd=12)
#intervalli di confidenza parametrici
segments(x,biblio_test.ci[,2],x,biblio_test.ci[,3],col="red",lwd=6)
#valori stimati
points(x,biblio_test.pi[,1],pch=19,col="cyan",cex=1.5)
points(biblio_test$spese_gestione,pch=19,col="yellow",cex=1.5)
```