Analisi delle Cause di Morte per Problemi Cardiaci

Progetto I per il corso di Statistica
Daniele Giaquinta
Corso di laurea in Artificial Intelligence and Data Engineering

Indice

1 Abstract

Lo scopo di questa relazione è studiare nei vari paesi la dipendenza tra le morti causate da problemi cardiaci e alcuni dati tra cui i più notoriamente correlati (obesità, consumo di sigarette, di alcol) e prendendo inizialmente in analisi anche il PIL procapite per scoprire se questo indice di ricchezza di un paese è correlato all'aumento delle abitudini da parte della popolazione dei fenomeni sopra citati, e conseguentemente anche alle morti stesse per problemi cardiaci. Si è partiti dunque da quattro tabelle di dati che contengono tali informazioni, queste sono state aggregate per poi tentare l'utilizzo di due modelli di regressione.

Essere in grado di comprendere l'impatto di queste abitudini sulla propria salute con dei riscontri numerici che ci informino meglio su quanto aumenti la probabilità di sviluppare patologie cardiache potrebbe contribuire a frenarle; per questo sarebbe tuttavia necessario ottenere un modello particolarmente accurato in modo da poterlo anche reputare affidabile.

2 Dati

Le tabelle di dati sono reperibili su https://ourworldindata.org/charts:

- **Percentuale di morti dovute a problemi cardiaci suddivise per anno e per stato** dalla tabella *Share of deaths from heart disease*; tale suddivisione è la stessa adottata anche dalle prossime tabelle pertanto non verrà ripetuta.
- **Percentuale di adulti in condizione di obesità** e **PIL pro-capite annuo di ogni stato** dalla tabella *Share of adults that are obese vs. GDP per capita.*
- **Consumo di alcol in litri per persona** con dati su individui che hanno almeno 15 anni di età dalla tabella *Alcohol consumption vs. GDP per capita* che ci fornisce inoltre la stessa informazione ridondante sul PIL e il numero di abitanti in ogni stato ogni anno.
- **Percentuale di fumatori abituali** sulla popolazione dalla tabella *Prevalence of daily smoking in populations.*

2.1 Aggregazione dei Dati

Per prima cosa le tabelle sono state unite usando come chiave comune lo stato e l'anno relativo ai dati raccolti. Sebbene alcune tabelle contenessero migliaia di tuple, avendo mantenuto dopo l'aggregazione solo tuple in cui non mancava nessuna delle informazioni sopra citate si è rimasti con 510 righe.

Per l'aggregazione sono state importate le tabelle in csv sulla workbench di *MySQL* e si è potuto unire le tabelle molto semplicemente grazie ad un *NATURAL INNER JOIN*, essendo le chiavi formattate in modo analogo in tutte le quattro tabelle.

2.2 Preprocessing e Correlazioni

I dati aggregati danno dunque origine ad una tabella formattata come segue Figura 1:

Si nota che il nome dello stato così come l'anno sono ininfluenti nella analisi regressiva, servono solo a riferire i dati, pertanto sono stati rimossi dalla tabella; non solo, anche il numero di abitanti di ogni stato è stato rimosso poiché i dati rappresentano informazioni percentuali, pertanto la popolazione non avrebbe alcuna influenza e correlazione con gli altri attributi.

Prima di discutere di regressione è necessaria una parentesi sull'attributo PIL pro-capite; come si può vedere dal diagramma di dispersione e dalla matrice di correlazioni Figura 2 l'ipotesi iniziale era corretta, ovvero esiste una correlazione tra il PIL pro-capite e le abitudini che influiscono sul rischio di problemi cardiaci (specialmente l'obesità), tuttavia l'effettiva correlazione lineare tra il PIL e l'incidenza stessa delle morti non è elevata e risulta che nel modello regressivo senza alcuna alterazione ai dati la presenza del PIL peggiora solo la proporzione di varianza spiegata dal

	Entity	Year	AlcoholConsumption	Population	DailySmoking	Obesity	GDPpercapita	HeartDiseaseDeaths
1	Afghanistan	2010	0.21	28189672	11.9	3.3	1957.0291	22.91
2	Albania	2000	6.57	3182027	20.0	12.8	5892.5903	45.71
3	Albania	2005	7.65	3032636	18.2	15.6	8040.0928	53.10
4	Albania	2010	7.69	2913402	19.8	18.7	10749.4814	55.31
5	Algeria	2000	0.58	30774624	15.8	15.0	8710.4443	41.23
6	Algeria	2005	0.81	32956690	12.9	17.9	10504.8447	43.08
7	Algeria	2010	0.65	35856348	10.9	21.4	10970.6924	44.42
8	Angola	2000	2.76	16394067	8.9	3.0	4727.9663	7.62
9	Angola	2005	4.89	19450962	8.9	3.9	6210.2222	8.41
10	Angola	2010	8.16	23364196	8.8	5.1	7692.4341	9.98
11	Antigua and Barbuda	2000	5.13	75070	3.6	12.1	18313.6289	35.18
12	Antigua and Barbuda	2005	5.37	79879	3.5	14.0	19683.8418	33.99
13	Antigua and Barbuda	2010	5.60	85710	3.4	16.1	18206.0156	33.19
14	Argentina	2000	8.75	37070772	23.7	20.5	18625.2832	32.80
15	Argentina	2005	8.66	39070504	23.2	22.8	19426.4395	30.84
16	Argentina	2010	9.09	41100124	20.5	25.3	23521.2695	29.65
17	Armenia	2000	4.23	3168525	26.7	14.4	4048.2500	51.01
18	Armenia	2005	5.79	3047254	26.3	15.7	7419.8711	50.85

Figura 1: Tabella iniziale dei dati

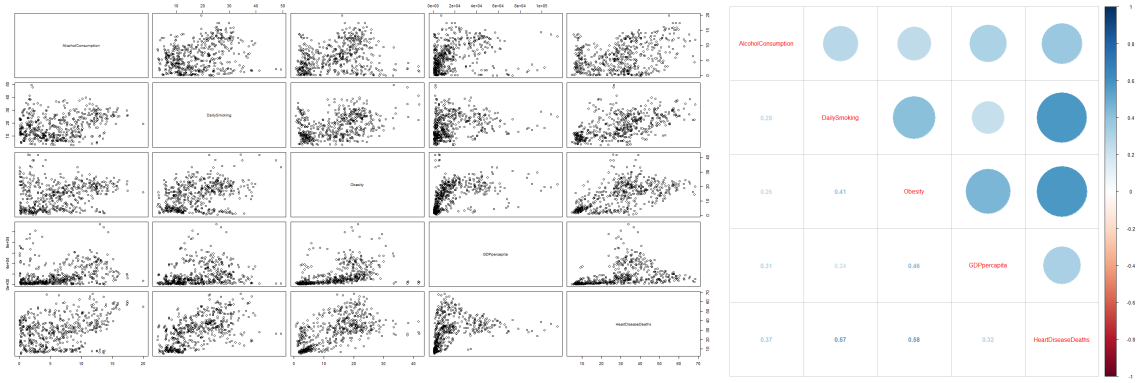


Figura 2: Diagramma di dispersione e matrice delle correlazioni

modello corretto lasciando inalterata quella non corretta; **nel modello lineare per tanto viene rimosso questo attributo** che come si osserverà tornerà rilevante nell'analisi non lineare.

Per quanto riguarda gli altri dati, tutti contribuiscono al modello, purtroppo si nota che le correlazioni non sono altissime seppur non indifferenti; nel modello ottimizzato si riusciranno a ricavare correlazioni leggermente più alte nell'ipotesi in cui le relazioni tra i dati non siano lineari.

3 Regressione Lineare

Per prima cosa è stato costruito il modello di regressione lineare con feature gli attributi *Obesity*, *AlcoholConsumption*, *DailySmoking* e *GDPpercapita*; sono stati poi rimossi attributi partendo dal *p-value* più grande Figura 3:

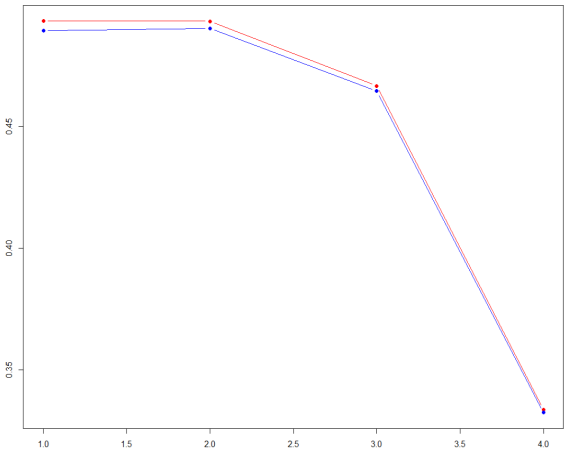
Si nota quanto già affermato in precedenza; il PIL non ha effetto sulla varianza (in rosso) ma addirittura diminuisce la varianza spiegata (in blu). Tra il punto 2 e il punto 3 invece si nota una variazione di pendenza non indifferente anche tenuta

```
Call:
lm(formula = HeartDiseaseDeaths ~ . - GDPpercapita, data = SecondTable)
Residuals:
    Min      1Q  Median      3Q     Max
-35.415  -7.418  -1.027   6.683  36.123

Coefficients:           Est.Std.      Error   t-value   Pr(>|t|)
(Intercept)              5.65021    1.20165     4.702   3.33e-06 ***
AlcoholConsumption       0.58523    0.11359     5.152   3.69e-07 ***
DailySmoking             0.62091    0.06039    10.282    < 2e-16 ***
Obesity                  0.66677    0.06123    10.889    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.71 on 506 degrees of freedom
Multiple R-squared:  0.4933, Adjusted R-squared:  0.4903
F-statistic: 164.2 on 3 and 506 DF,  p-value: < 2.2e-16
```

(a) Sommario del modello scelto



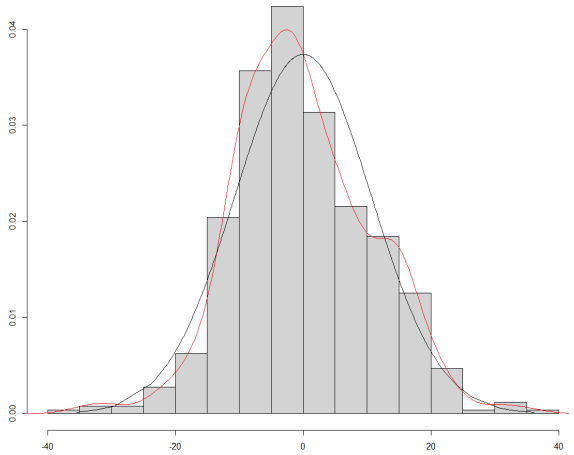
(b) Varianze nei 4 modelli

Figura 3

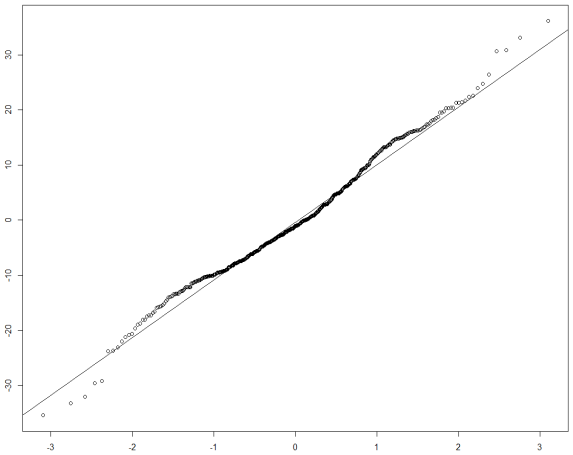
in considerazione la scala; **il modello scelto è pertanto il secondo da sinistra** il cui *p-value* è per altro molto piccolo.

3.1 Analisi dei Residui

Dopo la scelta del modello sono stati analizzati i residui. Segue l'istogramma per l'aderenza dei residui ad una densità gaussiana e il QQ plot Figura 4:



(a) Istogramma della normalità dei residui



(b) Plot Quantile-Quantile

Figura 4

In entrambi i casi si nota che l'aderenza al riferimento di normalità sembra buono, per avere una risposta più definitiva sulla gaussianità dei residui sono stati calcolati gli indici skewness e kurtosi ed è stato effettuato il Test di Shapiro-Wilk; mentre i primi due producono un buon risultato, quest'ultimo suggerisce che i residui non seguono il comportamento di una gaussiana normale in quanto lo score ottenuto è

inferiore a 0.05 Figura 5. Con questo modello lineare forse non si è quindi riusciti a catturare tutte le dipendenze e relazioni tra le variabili. Nemmeno con uno degli altri modelli lineari creati è risultato migliore lo score nello Shapiro Test; in realtà il risultato non è tragico poiché il test è stato effettuato sull'intero dataset di residui partendo da una quantità di dati non indifferente e il punteggio è lontano solo di due ordini di grandezza dal risultato sperato; come si vedrà, il modello non lineare riesce a risolvere questo probema.

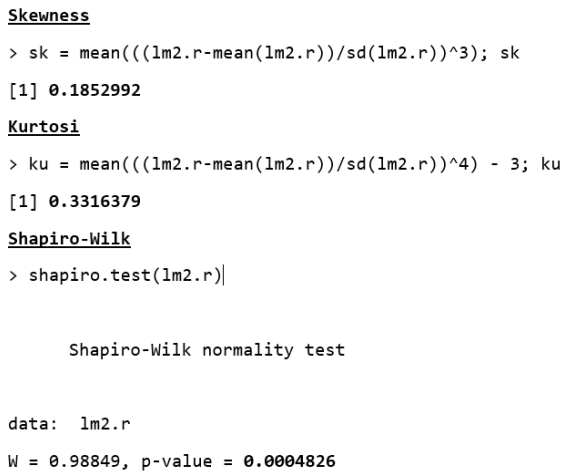


Figura 5: Indici Skewness, Kurtosi e Shapiro-Wilk Test

4 Regressione non Lineare

L'individuazione delle relazioni non lineari tra le variabili non è semplice a causa della dispersione dei dati che si può osservare nel relativo grafico. Giudicando dal grafico di dispersione e dalla matrice delle correlazioni, la trasformazione più efficace si è rivelata quella logaritmica per tutti gli attributi. Anche questa volta sono stati prodotti quattro modelli; è interessante osservare che il PIL pro-capite, attributo scartato nella regressione lineare, assume ora una correlazione con l'attributo target molto più alta.

La varianza e la varianza spiegata sono aumentate sensibilmente arrivando ad un punteggio finalmente accettabile, d'altronde non è ragionevole aspettarsi una regressione ottima con questi dati per motivi che saranno discussi nelle conclusioni Figura 6.

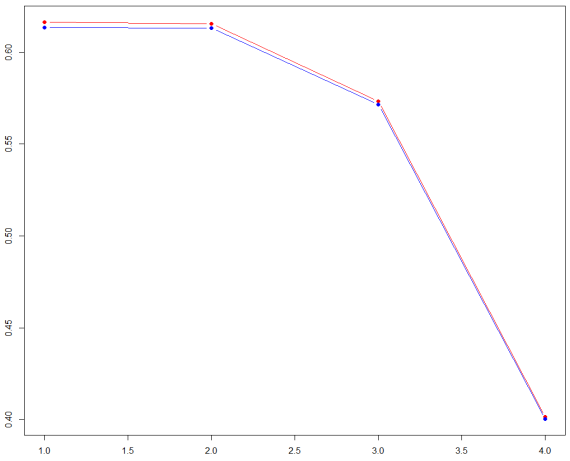
In questo caso il consumo di alcol per stato, con il p-value minore, è stato rimosso dal modello poiché i punteggi aumentavano solo leggermente e i residui risultavano più lontani dalla somiglianza ad una distribuzione normale; **è stato scelto dunque il secondo modello**.

```
Call:
lm(formula = HeartDiseaseDeaths ~ . - AlcoholConsumption, data = ThirdTable)
Residuals:
     Min       1Q   Median       3Q      Max
-1.03619 -0.27467  0.00382  0.26851  1.04901

Coefficients:  Estimate  Std.Error   t-value   Pr(>|t|)
(Intercept)    0.10689     0.15044     0.710     0.478
DailySmoking   0.43724     0.03404    12.844   < 2e-16 ***
Obesity        0.23733     0.02738     8.668   < 2e-16 ***
GDPpercapita   0.14916     0.02002     7.452  4.01e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3911 on 506 degrees of freedom
Multiple R-squared:  0.6154,  Adjusted R-squared:  0.6131
F-statistic: 269.8 on 3 and 506 DF,  p-value: < 2.2e-16
```

(a) Sommario del modello scelto

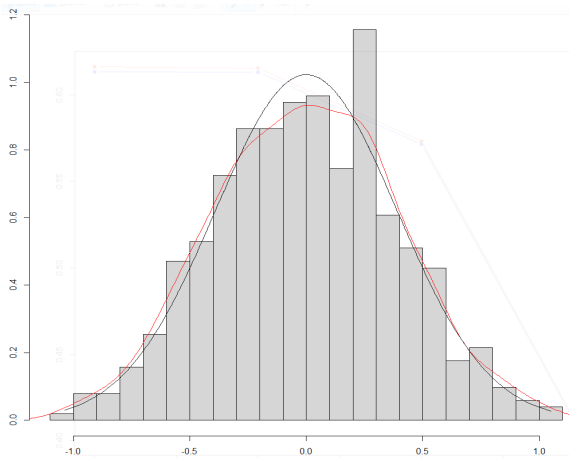


(b) Varianze nei 4 modelli

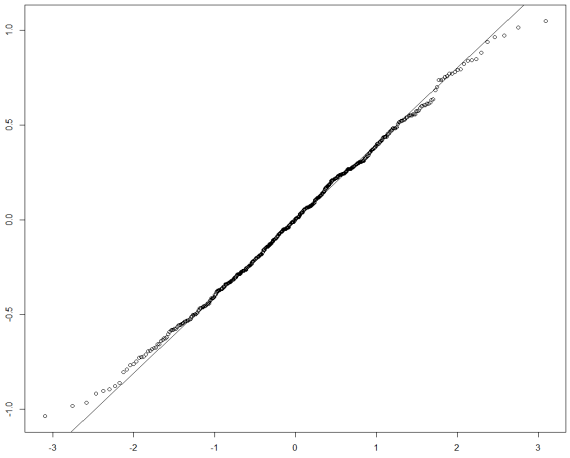
Figura 6

4.1 Analisi dei Residui

Le analisi effettuate sui residui sono le medesime Figura 7:



(a) Istogramma della normalità dei residui



(b) Plot Quantile-Quantile

Figura 7

Si nota dall'istogramma e dal diagramma Quantile-Quantile che l'aderenza dei residui al comportamento di una gaussiana normale sembra molto migliorato rispetto al modello lineare; se ne ha conferma definitiva con i test parametrici Figura 8 in cui non solo il punteggio nel test di skewness e Kurtosi rimangono accettabili, ma **risulta ottimo il punteggio ottenuto al test di Shapiro-Wilk** sull'intero insieme dei residui, si può quindi non rigettare l'ipotesi di normalità; è presumibile esser riusciti ad aver catturato in gran parte le relazioni tra i dati e l'attributo target.

```
Skewness
> sk = mean(((nlm.r-mean(nlm.r))/sd(nlm.r))^3); sk
[1] 0.01600756
Kurtosi
> ku = mean(((nlm.r-mean(nlm.r))/sd(nlm.r))^4) - 3; ku
[1] -0.313527
Shapiro-Wilk
> shapiro.test(nlm.r)

      Shapiro-Wilk normality test

data:  nlm.r
W = 0.99736, p-value = 0.5962
```

Figura 8: Indici Skewness, Kurtosi e Shapiro-Wilk Test

5 Predizione e Confronto dei Modelli

Al fine di confrontare i modelli sono stati divisi i dati in un insieme di training e uno di testing (470 e 40), sono stati confrontati nuovamente gli R-Squared ed è stato confrontato l'errore; a tal fine la suddivisione in sample randomici di dati di training e di testing è stata ripetuta 50 volte per avere risultati più distribuiti e tentare di osservare una tendenza. I risultati sono:

	Modello Lineare	Modello Logaritmico
R-Squared	0.4740726	0.6062671
Errore Medio	10.55922	10.33614
Deviazione Standard dell'Errore	1.432842	1.294754

Si può osservare l'andamento dell'errore nelle simulazioni in Figura 9; dal grafico ma soprattutto dalla tabella riportata sopra si nota che gli errori hanno un comportamento molto simile, sia a livello di andamento che di media che di deviazione standard. **Il modello migliore tra i due è quindi quello logaritmico** che seppur a parità di errore risulta più preciso in quanto a precisione nelle previsioni ma soprattutto presenta nei residui una piena tendenza alla normalità.

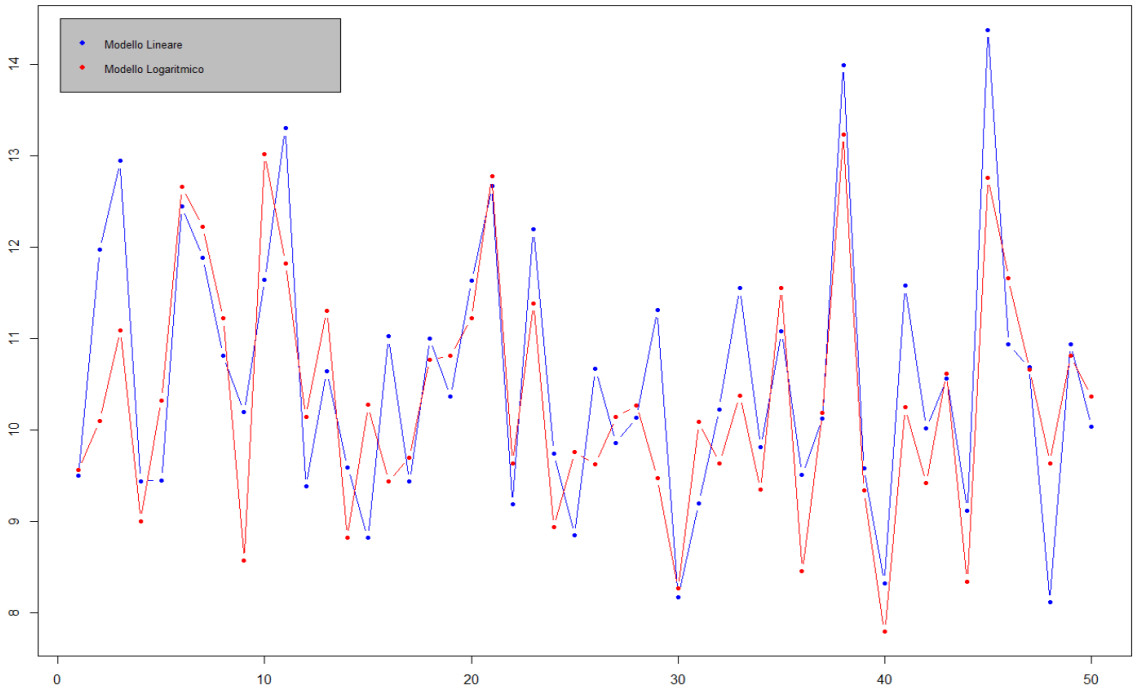


Figura 9: Confronto dell'errore nei due modelli

6 Conclusioni

Riassumendo, il modello migliore risulta essere quello logaritmico e oltre alla differenza nella varianza spiegata si nota la differenza di importanza dedicata agli attributi della regressione; nel modello lineare si poteva scartare il PIL pro-capite mentre in quello logaritmico si può scartare il consumo di alcol medio per persona. Questo si potrebbe spiegare pensando che insorgono problemi cardiaci solo in seguito a situazioni di alcolismo abituale (e fuori controllo), ma un valore superiore di consumo di alcol medio potrebbe riflettersi non necessariamente in questo, ma anche solo in una maggiore percentuale di popolazione che consuma alcol e questo dipende anche dalle culture dei paesi in analisi. Il PIL invece è un indice di ricchezza con una correlazione maggiore a tutte le abitudini negative che possono portare a problemi cardiaci; non solo, paesi con PIL pro-capite alto tendono ad avere una maggiore speranza di vita e con l'aumento dell'età aumenta anche il rischio di tali patologie.

Il modello ottenuto è ritenuto soddisfacente e il punteggio non altissimo ottenuto nella regressione è facilmente attribuibile alla mancanza di altri dati notoriamente legati all'insorgere di problemi cardiaci quali la precedentemente citata speranza di vita, lo stress, l'incidenza del diabete o la predisposizione genetica; gli ultimi soprattutto rappresentano dati di cui non è stato possibile reperire dati sufficienti che avrebbero potuto migliorare ulteriormente il modello.

Appendice

Segue il codice in R dell'analisi:

```
install.packages('nortest')
library(nortest)
library(corrplot)

#Tabella iniziale che contiene anche la popolazione priva di
   alcuna correlazione col resto
#plot(FirstTable)
#corrplot(cor(FirstTable), 'number')

#IMPORTARE QUI IL FILE tabella.csv
SecondTable = read.csv("tabella.csv")
plot(SecondTable)
corrplot.mixed(cor(SecondTable))


#REGRESSIONE LINEARE SU 4 POSSIBILI MODELLI
r=matrix(ncol=2,nrow=4)                  #memorizzo la varianza
    e la varianza spiegata nella matrice "r"
lm = lm(HeartDiseaseDeaths ~ ., data = SecondTable)
summary(lm)
r[1,]=c(summary(lm)$r.squared, summary(lm)$adj.r.squared)
#modello senza il PIL pro-capite
lm = lm(HeartDiseaseDeaths ~ .-GDPpercapita, data =
   SecondTable) #QUESTO E' IL MODELLO SCELTO
summary(lm)
r[2,]=c(summary(lm)$r.squared, summary(lm)$adj.r.squared)
#modello senza consumo di alcol per stato
lm = lm(HeartDiseaseDeaths ~ .-GDPpercapita-AlcoholConsumption
   , data = SecondTable)
summary(lm)
r[3,]=c(summary(lm)$r.squared, summary(lm)$adj.r.squared)
#modello senza percentuale di fumatori abituali
lm = lm(HeartDiseaseDeaths ~ .-GDPpercapita-AlcoholConsumption
   -DailySmoking, data = SecondTable)
summary(lm)
r[4,]=c(summary(lm)$r.squared, summary(lm)$adj.r.squared)

#ESAMINO IL CALO DI VARIANZA E VARIANZA SPIEGATA
ymin=min(r)
ymax=max(r)
plot(r[,1],pch=19,type="b",col="red",ylim=c(ymin,ymax))
lines(r[,2],pch=19,type="b",col="blue")


#ANALISI DEI RESIDUI
lm.r=residuals(lm)
plot(fitted(lm),lm.r,pch=20)
```

```
#Esaminiamo la possibile aderenza dei residui a una
   distribuzione Gaussiana attraverso  l osservazione  della
   densit  empirica
hist(lm.r,20,freq=F)
lines(density(lm.r),col="red")
m=mean(lm.r)
s=sd(lm.r)
lines(sort(lm.r),dnorm(sort(lm.r),m,s))
#distribuzione empirica
plot(ecdf(lm.r),pch=".")
y=seq(m-3*s,m+3*s,6*s/100)
lines(y,pnorm(y,m,s),col="red")
#quantile-quantile plot
qqnorm(lm.r)
qqline(lm.r)
#skewness
sk = mean(((lm.r-mean(lm.r))/sd(lm.r))^3); sk
#kurtosi
ku = mean(((lm.r-mean(lm.r))/sd(lm.r))^4) - 3; ku
#I residui non hanno proprio distribuzone gaussiana quindi
   shapiro test:
shapiro.test(lm.r)
```

```
#CREAZIONE DELLA TABELLA PER LA REGRESSIONE NON LINEARE (
   LOGARITMICA)
ThirdTable = log(SecondTable)
plot(ThirdTable)
corrplot.mixed(cor(ThirdTable))

#REGRESSIONE NON LINEARE SU 4 POSSIBILI MODELLI
r=matrix(ncol=2,nrow=4)                 #memorizzo la varianza
    e la varianza spiegata nella matrice "r"
nlm = lm(HeartDiseaseDeaths ~ ., data = ThirdTable)
summary(nlm)
r[1,]=c(summary(nlm)$r.squared, summary(nlm)$adj.r.squared)
#modello senza consumo di alcol
nlm = lm(HeartDiseaseDeaths ~ .-AlcoholConsumption, data =
   ThirdTable) #QUESTO E' IL MODELLO SCELTO
summary(nlm)
r[2,]=c(summary(nlm)$r.squared, summary(nlm)$adj.r.squared)
#modello senza PIL pro-capite
nlm = lm(HeartDiseaseDeaths ~ .-GDPpercapita-
   AlcoholConsumption, data = ThirdTable)
summary(nlm)
r[3,]=c(summary(nlm)$r.squared, summary(nlm)$adj.r.squared)
#modello senza percentuale di fumatori abituali
nlm = lm(HeartDiseaseDeaths ~ .-GDPpercapita-
   AlcoholConsumption-DailySmoking, data = ThirdTable)
```

```
summary(nlm)
r[4,]=c(summary(nlm)$r.squared, summary(nlm)$adj.r.squared)
#ESAMINO IL CALO DI VARIANZA E VARIANZA SPIEGATA
ymin=min(r)
ymax=max(r)
plot(r[,1],pch=19,type="b",col="red",ylim=c(ymin,ymax))
lines(r[,2],pch=19,type="b",col="blue")


#ANALISI DEI RESIDUI
nlm.r=residuals(nlm)
plot(fitted(nlm),nlm.r,pch=20)
#Esaminiamo la possibile aderenza dei residui a una
   distribuzione Gaussiana attraverso  l osservazione  della
   densit  empirica
hist(nlm.r,20,freq=F)
lines(density(nlm.r),col="red")
m=mean(nlm.r)
s=sd(nlm.r)
lines(sort(nlm.r),dnorm(sort(nlm.r),m,s))
#distribuzione empirica
plot(ecdf(nlm.r),pch=".")
y=seq(m-3*s,m+3*s,6*s/100)
lines(y,pnorm(y,m,s),col="red")
#quantile-quantile plot
qqnorm(nlm.r)
qqline(nlm.r)
#skewness
sk = mean(((nlm.r-mean(nlm.r))/sd(nlm.r))^3); sk
#kurtosi
ku = mean(((nlm.r-mean(nlm.r))/sd(nlm.r))^4) - 3; ku
#I residui non hanno proprio distribuzone gaussiana quindi
   shapiro test:
shapiro.test(nlm.r)
ks.test(nlm.r, "pnorm")
ad.test(nlm.r)



#CONFRONTO DEI MODELLI IN PREVISIONE
#divisione in dati di training e di testing
testset = sort(sample(470,40))
lm_train = SecondTable[-testset,]
lm_test = SecondTable[testset,]
nlm_train = ThirdTable[-testset,]
nlm_test = ThirdTable[testset,]
lm_train.lm=lm(HeartDiseaseDeaths~.-GDPpercapita,data=lm_train
   )
nlm_train.lm=lm(HeartDiseaseDeaths~.-AlcoholConsumption,data=
   nlm_train)
summary(lm_train.lm)$r.squared
```

```
summary(nlm_train.lm)$r.squared
#creazione delle variabili di predizione
lm_train.lm.p = predict(lm_train.lm,lm_test)
nlm_train.lm.p = predict(nlm_train.lm,nlm_test)
#diagramma delle previsioni dei due modelli
gmin=min(lm_train.lm.p,exp(nlm_train.lm.p),lm_test$
   HeartDiseaseDeaths)
gmax=max(lm_train.lm.p,exp(nlm_train.lm.p),lm_test$
   HeartDiseaseDeaths)
plot(lm_test$HeartDiseaseDeaths,pch=20,ylim=c(gmin,gmax))
points(lm_train.lm.p,col="blue",pch=20)
points(exp(nlm_train.lm.p),col="red",pch=20)
legend("topleft",inset=0.02,c("dati","modello lineare","
   modello logaritmico"),col=c("black","blue","red"),pch=c
   (19,19),bg="gray",cex=.8)

#comparazione degli errori di previsione
n=50
err_lin = rep(0,n)
err_log = rep(0,n)
for(i in 1:n){
  testset <- sort(sample(470,40))
  lm_train <- SecondTable[-testset,]
  lm_test <- SecondTable[testset,]
  nlm_train <- ThirdTable[-testset,]
  nlm_test <- ThirdTable[testset,]
  lm_train.lm <- lm(HeartDiseaseDeaths~.-GDPpercapita,data=lm_
   train)
  nlm_train.lm <-lm(HeartDiseaseDeaths~.-AlcoholConsumption,
   data=nlm_train)

  lm_train.lm.p = predict(lm_train.lm,lm_test)
  nlm_train.lm.p = predict(nlm_train.lm,nlm_test)

  err_lin[i]=sqrt(mean((lm_train.lm.p - lm_test$
   HeartDiseaseDeaths)^2))
  err_log[i]=sqrt(mean((exp(nlm_train.lm.p) - lm_test$
   HeartDiseaseDeaths)^2))
}
mean(err_lin)
mean(err_log)
sd(err_lin)
sd(err_log)
gmin=min(err_lin,err_log)
gmax=max(err_lin,err_log)
plot(err_lin,type="b",pch=20,col="blue",ylim=c(gmin,gmax))
points(err_log,type="b",pch=20,col="red")
legend("topleft",inset=0.02,c("Modello Lineare","Modello
   Logaritmico"),col=c("blue","red"),pch=c(19,19),bg="gray",
   cex=.8)
```