Report Tecniche di Indagine Statistica: Analisi dei Microdati

Barrasso Marco, Giacomini Elisa, Guerriero Enrico, Insaghi Edoardo, Suklan Andrea
Secondo appello, sessione invernale 2023

Abstract

Il dataset AVQ dell'Istat (Aspetti della Vita Quotidiana) fa parte di un sistema integrato di indagini sociali, le indagini Multiscopo sulle famiglie, e rileva le informazioni fondamentali relative alla vita quotidiana degli individui e delle famiglie.
L'obiettivo dell'analisi è individuare, tra le variabili del dataset, una variabile Y di risposta e un insieme di variabili esplicative che descrivano la variabile risposta all'interno di un modello logistico.

1 Il database

Il database AVQ dell'Istat è il risultato di un sondaggio condotto su 20.000 famiglie e circa 50.000 individui e conta 755 variabili.
Prima di scegliere una variabile risposta e le seguenti variabili esplicative, è stato necessario fare una prima selezione, basata sulle risposte mancanti.
Infatti, contando il sondaggio di molte domande, gli intervistati si sono astenuti da parte di queste, rendendo il database pieno di NA.
La maggiorparte delle variabili sono qualitative o quantitative divise in classi, pertanto alcune variabili divise in classi sono state considerate quantitative.

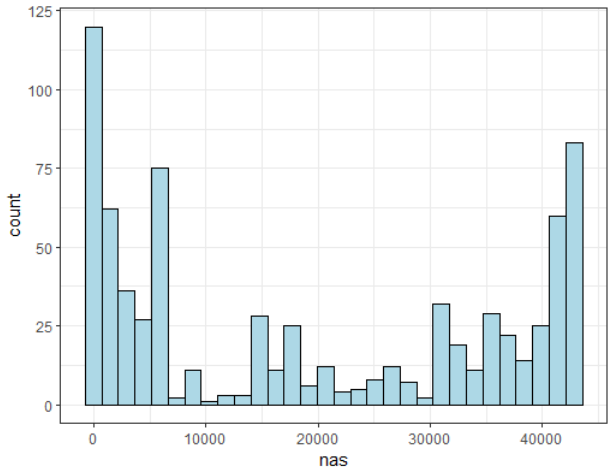
Pulizia del database

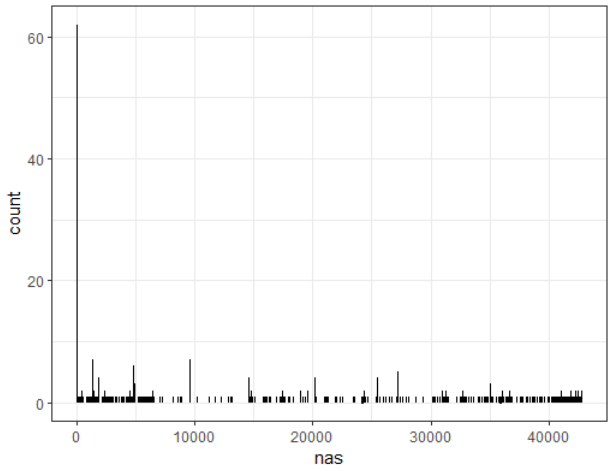
Prima di proporre la lista delle variabili, si effettua una scrematura sulla base delle risposte non date.
La prima scrematura avviene mediante un ciclo for che itera tutte le colonne del database e inserisce in un vettore, inizializzato precedentemente, la somma degli NA conteggiati per ogni colonna:

```
nas <- c()
for (i in 1:length(data[1,])){
    nas <- c(nas, sum(is.na(data[,i])))
    }
```

Graficamente, è possibile rappresentare gli NA con un istogramma:

```
ggplot(aes(x=nas), data=data.frame("nas"=nas)) +
    geom_histogram(col="black", fill="lightblue") + theme_bw()
ggplot(aes(x=nas), data=data.frame("nas"=nas)) +
    geom_bar(col="black", fill="blue") + theme_bw()
```





Si può osservare dai grafici come ci sono molte colonne con 0 NA, ma allo stesso modo ce ne sono molte che hanno significativamente molti NA.
Possiamo anche osservare qual è la percentuale di colonne che presentano meno di 100 NA:

```
mean(nas < 100)*100
```

[1] 14.43709

Diventa pertanto fondamentale selezionare una soglia massima oltre la quale gli NA di una colonna vengono considerati "troppi".
La prima cosa da fare è fissare una soglia, indicante il numero di NA presenti in una colonna, oltre la quale si scarta la colonna. Bisogna anche tenere a mente che, una volta fissato il numero di colonne, tutte le righe contenenti almeno un NA verranno scartate. Quindi, più saranno le colonne mantenute nel database, meno saranno le righe disponibili per la costruzione del modello.

Per avere un'idea, sono state fissate 3 soglie diverse:

- 100: 115 variabili e 42719 righe
- 1000: 130 variabili e 37781 righe
- 10000: 333 variabili e 9263 righe

Ricordando che il database completo conta 755 variabili 42810 righe.

Saranno quindi scartate le colonne che superano la soglia di NA fissata a 1750. La soglia è una soglia selezionata empiricamente, cercando di bilanciare il più possibile tra numero di variabili scartate e numero di righe perse poiché contenenti NA.

```
lim <- 1750
data <- data[, nas<lim]
#remove rows that have one or more na
data <- na.omit(data)
```

Ora si visualizzano quante variabili sono rimaste da questa selezione.

```
#cols remaining
length(data[1,])
```

[1] 176

```
#rows remaining
length(data[,1])
```

[1] 25437

A questo punto ha senso visualizzare le etichette delle colonne rimaste dopo questa prima selezione:[1]

```
names(data)
```

Variabile risposta: fumo

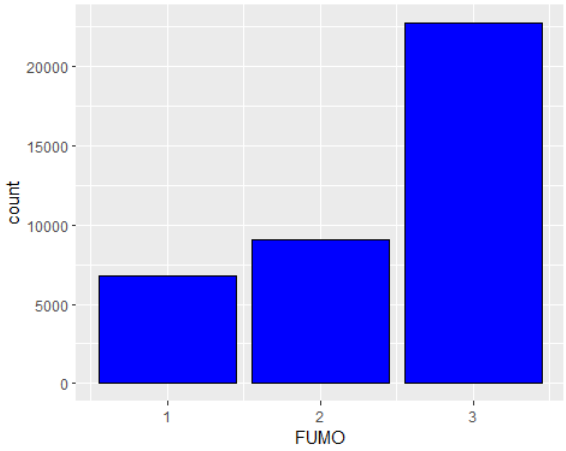
Ora il database appena definito è sufficientemente accurato e con un numero sufficientemente grande di variabili per poter scegliere una variabile risposta.
La prima variabile risposta scelta è "FUMO", che assume i seguenti valori:

- 1: la persona fuma
- 2: la persona ha mai fumato in passato
- 3: la persona non ha mai fumato

[1] Per visualizzare l'output del seguente comando, vedere la tabella 1

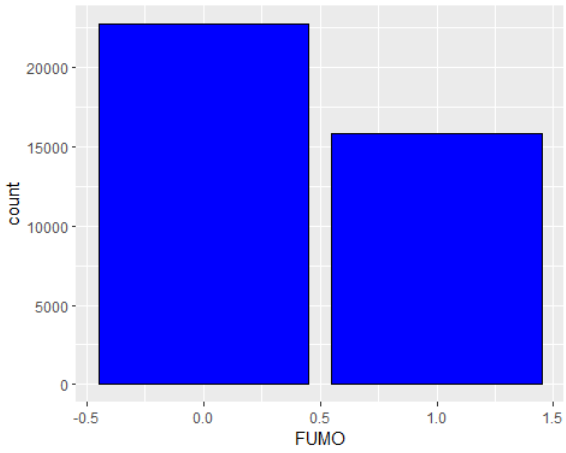
Il primo problema che si presenta è la non - dicotomicità della variabile; bisogna decidere se considerare gli ex fumatori tra i fumatori o tra i non fumatori.
Per logica bisognerebbe mettere gli ex fumatori tra i non fumatori, ma prima è bene visualizzare con chiarezza il database.
Per prima cosa è utile stampare un istogramma con la distribuzione dei 3 valori:

```
ggplot(aes(x=FUMO), data=data) +
geom_bar(col="black", fill="blue")
```



Nonostante quanto detto in precedenza, alla luce del database gli ex fumatori vengono accorpati ai fumatori. Questo può creare delle distorsioni rispetto ad un'analisi in cui li si accorpa con i non fumatori, ma in realtà l'analisi risulta comunque valida poiché si può cercare una correlazione con le cause che spingono una persona a fumare.

```
for (i in 1:length(data$FUMO)){
  if(! is.na(data$FUMO[i])){
    if(data$FUMO[i]==2 | data$FUMO[i]==1){
      data$FUMO[i] <- 1
    }
    if(data$FUMO[i]==3){
      data$FUMO[i] <- 0
    }
  }
}
```



Il risultato di questa variazione è che la variabile ottenuta è una variabile che assume due valori:

- 0: la persona non ha mai fumato
- 1: la persona è un fumatore o un ex fumatore.

Una seconda pulizia

Identificata la variabile risposta, è necessario selezionare le variabili esplicative. Ormai diventa necessario guardare la descrizione di tutte le variabili, una ad una, e scartare quelle che non sembrano pertinenti.
Dalla selezione "manuale" restano le variabili contenute nel vettore keep:

```
keep <- c("NCOMP", "ETAMi", "SESSO", "REGMF", "CITTMi", "SALUTE", "SENELE",
          "SPORCO", "INQAR", "RUMORE", "CRIM", "PARCHI", "STANZEM",
          "GODAB", "AVVOC", "BIC", "LIBFAM", "SITE", "RISEC", "FUMO")
```

Una breve analisi di ciascuna variabile rimasta.[2] Le variabili elencate nella tabella sono le uniche che vengono tenute nel database:

```
data <- data[, keep]
head(data)
keep %in% names(data)
```

Come ultima cosa, nella preparazione del database, è necessario fattorizzare le variabili qualitative:

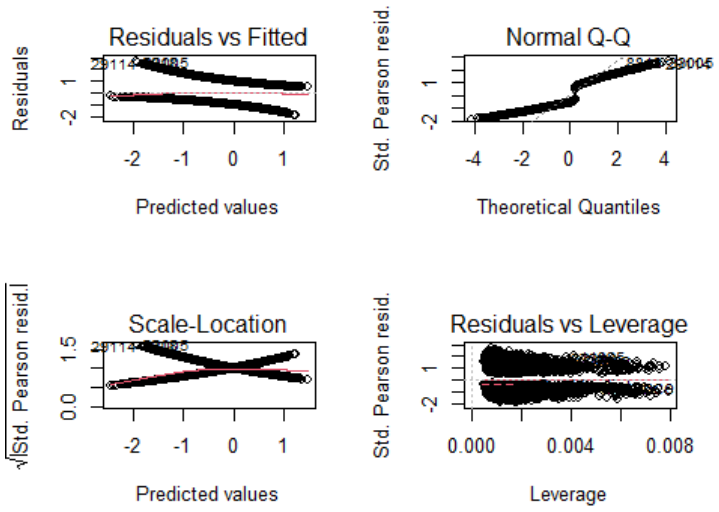
```
data$SESSO <- factor(data$SESSO)
data$REGMF <- factor(data$REGMF)
data$CITTMi <- factor(data$CITTMi)
data$PARCHI <- factor(data$PARCHI)
data$BIC <- factor(data$BIC)
data$GODAB <- factor(data$GODAB)
```

2 Il modello logit

Il database è pronto, con le sole variabili plausibilmente significative rispetto al fenomeno. Adesso si può fare un primo tentativo di regressione e valutare la significatività effettiva nel modello.[3]

```
model <- glm(FUMO ~ ., data=data, family=binomial())
summary(model)


par(mfrow=c(2,2))
plot(model)
```



[2] Tabella 2 a fine PDF
[3] L'output del seguente codice si trova nella tabella 3 a fine PDF

Molte variabili non risultano significative, e i residui non sono accettabili.
Prima di rimuovere manualmente le variabili non significative si effettua uno step AIC:

```
model1 <- stepAIC(model, direction="both", data=data)
summary(model1)
```

Il risultato dello step AIC[4] porta a questo modello:[5]

```
summary(model1)
```

Con lo step AIC si è ridotto notevolmente il numero di variabili, ma ancora molte sono non significative.
Per arrivare al modello finale si tolgono al modello ottenuto le variabili poco significative:[6]

```
modello3 <- glm(FUMO ~ ETAMi + AVVOC + SESSO + RISEC + SENELE + SPORCO + LIBFAM,
    family=binomial(), data=data)
summary(modello3)
```

Il modello definitivo, quindi, è in funzione delle variabili:

- Età in anni compiuti
- Negli ultimi 12 mesi la famiglia è ricorsa alla consulenza di un avvocato
- Sesso
- Come sono state complessivamente le risorse economiche della famiglia negli ultimi 12 mesi
- Soddisfazione complessiva del servizio dell'energia elettrica
- Presenza di sporcizia nelle strade nella zona in cui abita
- Quanti libri possiede la famiglia

Un modello alternativo

Ora si analizza un metodo alternativo per arrivare ad un modello. Partendo dal database ripulito, prima di effettuare lo step AIC, le variabili vengono selezionate secondo un altro criterio: si tengono solo le variabili quantitative o le variabili categoriche che possono però essere trattate come tali. Le variabili di interesse sono contenute nel vettore keep2:

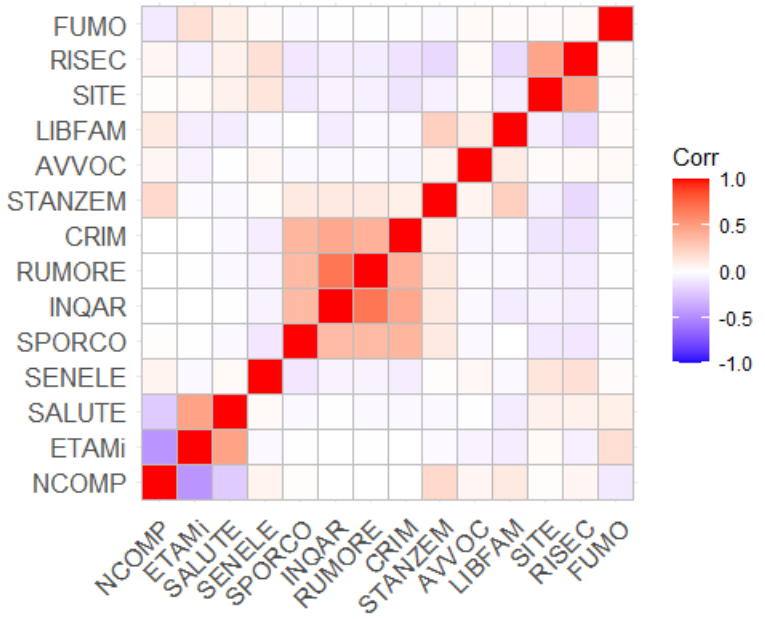
```
# quantitative variables
keep2 <- c("NCOMP", "ETAMi", "SALUTE", "SENELE", "SPORCO", "INQAR", "RUMORE",
           "CRIM", "STANZEM", "AVVOC", "LIBFAM", "SITE", "RISEC", "FUMO")
df <- data[, keep2]
head(df)
```

NCOMP	ETAMi	SALUTE	SENELE	SPORCO	INQAR	RUMORE	CRIM	STANZEM	AVVOC	LIBFAM	SITE	RISEC	FUMO
1	4	9	2	2	4	2	3	3	3	1	5	4	3	0
4	4	9	2	2	4	2	3	3	3	1	5	4	3	1
5	3	13	3	2	4	2	2	3	5	1	8	3	2	0
6	3	13	3	2	4	2	2	3	5	1	8	3	2	0
7	3	9	2	2	4	2	2	3	5	1	8	3	2	0
8	3	11	2	2	4	2	2	2	4	1	7	4	3	0

Il vantaggio di aver selezionato le variabili quantitative è che permette di effettuare dei confronti migliori per quanto riguarda la correlazione tra variabili:

[4] Ultimo passaggio dello step AIC nella tabella 4
[5] L'output del seguente codice è riportato nella tabella 5
[6] L'output del seguente codice è riportato nella tabella 6

```
ggcorrplot(cor(df))
```



Il grafico mostra le correlazioni tra le varie variabili, e si può osservare come la variabile FUMO abbia poche relazioni con le altre variabili:

- Ha correlazione negativa con il numero di componenti della famiglia, che è interpretabile come:
  Più la famiglia è numerosa, meno i componenti della famiglia tendono a fumare.
- Ha correlazione positiva con l'età dell'intervistato, interpretabile come:
  Più le persone sono grandi, più tendono a fumare.
- Ha correlazione positiva con la salute, interpretabile come:
  Meno le persone sono in salute, più tendono a fumare (la variabile salute, come verrà spiegato in seguito, ha i valori "al contrario").

Nel paragrafo delle conclusioni seguiranno considerazioni su questi risultati.

3 Analisi delle componenti principali

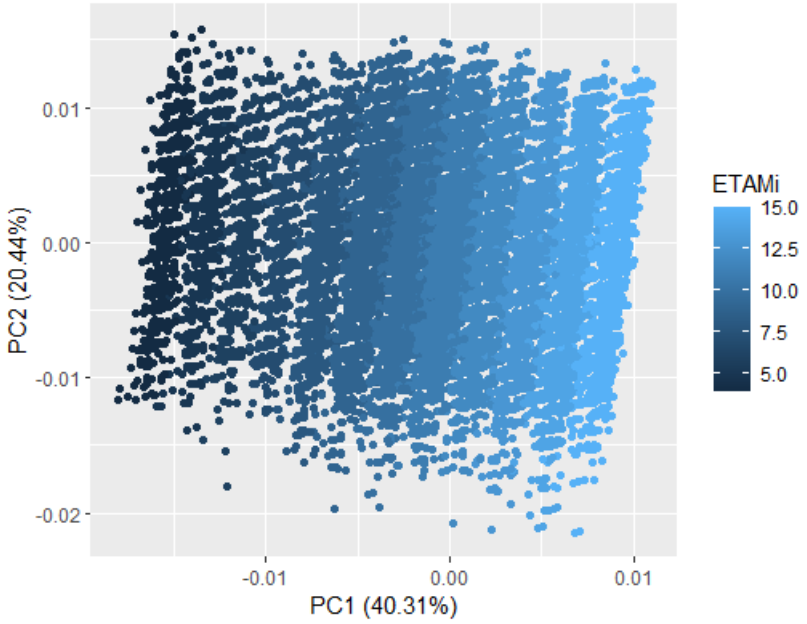
Per effettuare un'analisi delle componenti principali, si esegue questo codice:

```
data.pca <- prcomp(df)
summary(data.pca)
```

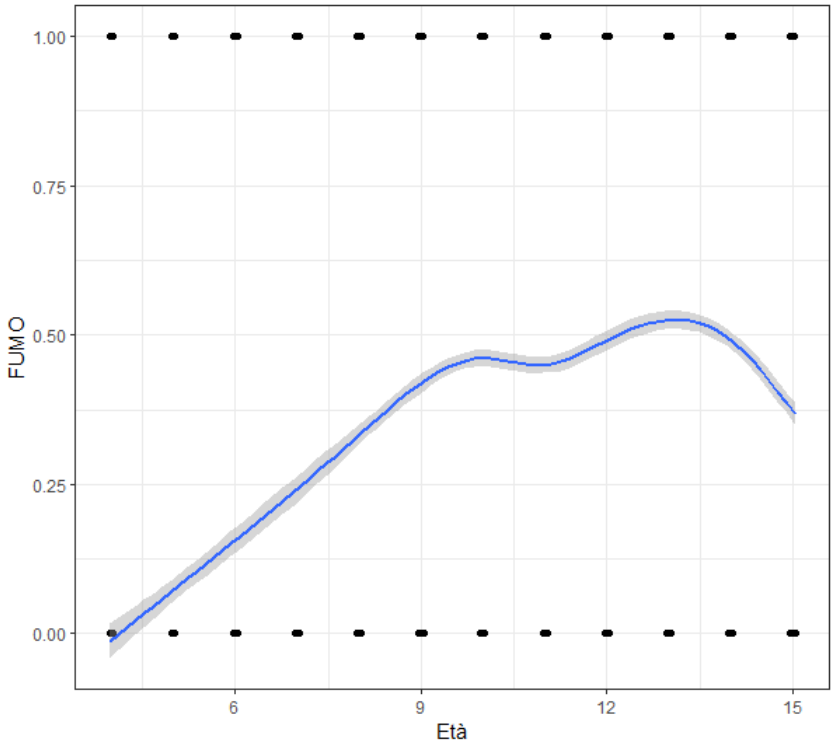
```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.9051 2.0687 1.5441 1.28494 1.02709 0.79240 0.73410 0.67334 0.63194
Proportion of Variance 0.4031 0.2044 0.1139 0.07887 0.05039 0.02999 0.02574 0.02166 0.01908
Cumulative Proportion  0.4031 0.6075 0.7214 0.80029 0.85068 0.88068 0.90642 0.92808 0.94715
                          PC10    PC11    PC12    PC13    PC14
Standard deviation     0.57563 0.49679 0.48443 0.42251 0.33915
Proportion of Variance 0.01583 0.01179 0.01121 0.00853 0.00549
Cumulative Proportion  0.96298 0.97477 0.98598 0.99451 1.00000
```

L'analisi delle componenti principali permette anche una visualizzazione grafica:

```
autoplot(data.pca, col="ETAMi", data=df)
```

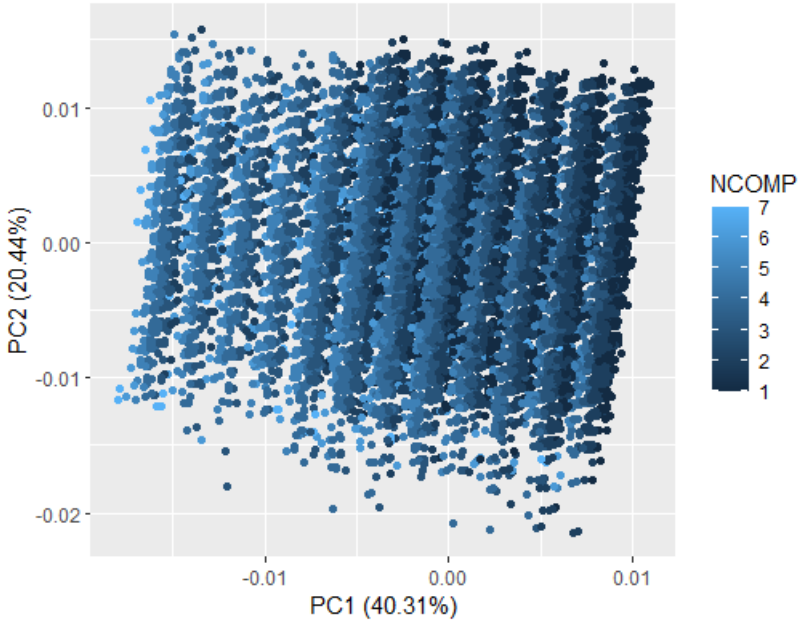


Da questo grafico emerge l'importanza della componente dell'età nel modello, la quale da sola spiega il 40% della varianza.
La correlazione tra età e fumo è rappresentabile anche con un grafico più leggibile:



L'altra variabile con una correlazione forte con il fumo è il numero di componenti della famiglia:

```
autoplot(data.pca, col="NCOMP", data=df)
```



Anche da questo grafico risulta che la variabile in questione, il numero di componenti della famiglia, spiega all'incirca il 40% della varianza. Tuttavia questo risultato non è particolarmente sorprendente, dal momento che si è osservata una forte correlazione tra numero di componenti della famiglia dell'intervistato e l'età dell'intervistato.

4 Un'altra variabile risposta: la criminalità

La variabile "criminalità" indica quanto viene percepita pericolosa la zona in cui si abita per l'intervistato. Le determinazioni possibili sono:

- 1: tanto pericolosa
- 2: abbastanza pericolosa
- 3: poco pericolosa
- 4: per niente pericolosa

Come prima cosa è necessario "dicotomizzare" la variabile: affinché funzioni il modello logit è necessario che abbia solo due determinazioni.
Come determinazioni sono state scelte le seguenti:

- 0: assenza di criminalità (per niente pericolosa o poco pericolosa)
- 1: presenza di criminalità (abbastanza pericolosa o tanto pericolosa)

```
newdata = data
newdata$CRIM = rep(0,length(data$CRIM))
newdata$CRIM[data$CRIM <= 2] = 1
```

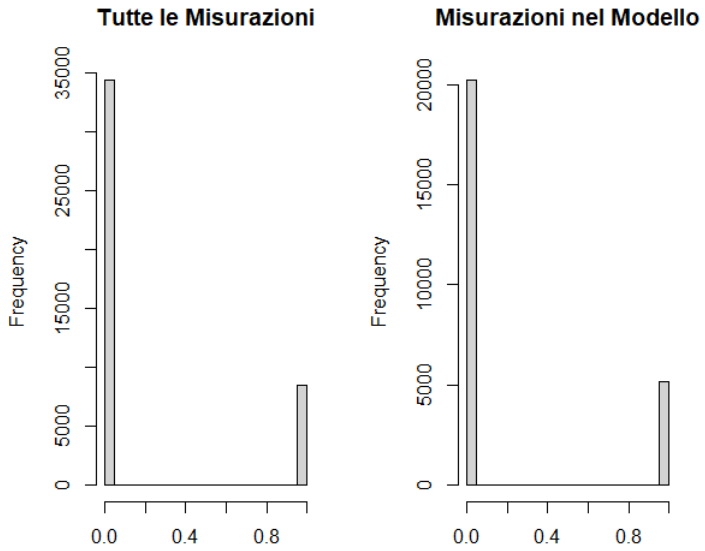
Le variabili che vengono prese in considerazione nella creazione del modello sulla criminalità sono:

```
names(newdata)
```

```
[1] "NCOMP"   "ETAMi"   "SESSO"   "REGMF"   "CITTMi"  "SALUTE"  "SENELE"  "SPORCO"  "INQAR"
[10] "RUMORE"  "CRIM"  "PARCHI"  "STANZEM" "GODAB"   "AVVOC"   "BIC"     "LIBFAM"  "SITE"
[19]  "RISEC"    "FUMO"
```

Dal momento che il database è stato pulito tenendo a mente di regredire rispetto alla variabile ”fumo”, prima di procedere è necessario controllare che le risposte nel database grezzo siano coerenti rispetto a quelle nel database pulito:

```
par(mfrow=c(1,2))
data.crim = data.pure[!is.na(data.pure$CRIM),]
sum(is.na(data.crim$CRIM))
data.crim.2 = data.crim
data.crim.2$CRIM = rep(0,length(data.crim$CRIM))
data.crim.2$CRIM[data.crim$CRIM <= 2] = 1
hist(data.crim.2$CRIM,main="Tutte le Misurazioni",xlab="")
hist(newdata$CRIM,main="Misurazioni nel Modello",xlab="")
```



Come si evince dai grafici, la differenza tra la variabile contenuta nel database pulito e quella nel database complessivo è irrisoria.

Costruzione del modello

Il primo modello prevede la regressione della criminalità su tutte le altre variabili[7]:

```
crim.mod1 = glm(CRIM ~ .,data = newdata,family=binomial())
summary(crim.mod1)
```

[7] L’output è riportato nella tabella 7

Dalla tabella si osserva come le variabili significative siano molte più rispetto al modello con il fumo, ma ce ne sono ancora diverse non significative.
Per alleggerire il modello si effettua come prima lo step AIC[8]:

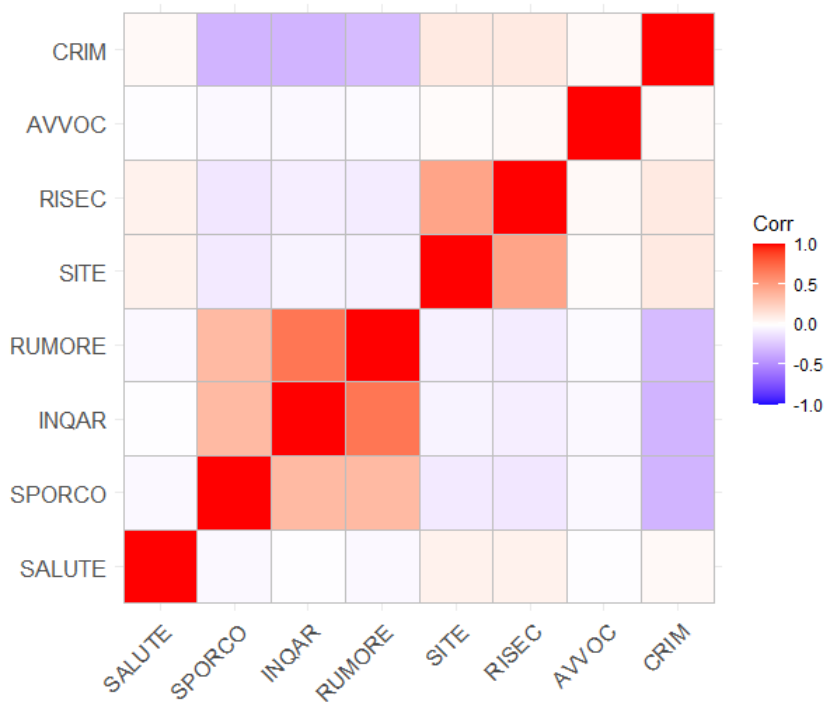
```
crim.aic = stepAIC(crim.mod1,direction="both",data=newdata)
summary(crim.aic)
```

Dal momento che sono presenti ancora molte variabili esplicative, per creare un modello più leggibile è necessario fare una selezione.
Il criterio utilizzato per alleggerire i dati è selezionare le variabili più influenti nel modello e le cui correlazioni potrebbero essere non del tutto spurie.
Nel vettore keep.crim sono contenute tutte le variabili mentre nel vettore keep.crim.df sono contenute le variabili quantitative con le quali verrà costruito il modello finale.

```
keep.crim = c("CITTMi","SALUTE","SPORCO","INQAR","RUMORE","PARCHI","GODAB",
    "SITE","RISEC","AVVOC","CRIM")
keep.crim.df = c("SALUTE","SPORCO","INQAR","RUMORE","SITE","RISEC","AVVOC","CRIM")
df.crim = newdata[,keep.crim.df]
```

Essendo tutte le variabili quantitative, è possibile stampare il correlogramma:

```
ggcorrplot(cor(df.crim))
```



[8]Gli output dell'ultimo passo dello step AIC e del summary sono nelle tabelle 8 e 9 a fine PDF

Da questo correlogramma emergono delle correlazioni evidenti:

- La criminalità è correlata positivamente, più o meno nella stessa maniera, con la sporcizia, l'inquinamento e il rumore. Nel grafico è segnata come una correlazione negativa, ma solamente perché le variabili "SPORCO", "INQAR" e "RUMORE" sono categoriche ordinate dove 1 rappresenta la massima intensità e 4 la minima, quindi sono "al contrario".
- La criminalità è correlata negativamente con le due variabili che riguardano il reddito e il benessere economico ("SITE" e "RISEC" assumono i valori "al contrario" come le altre)
- La criminalità è pressoché incorrelata con la salute e l'essersi rivolti ad un avvocato negli ultimi 12 mesi.

Il modello finale si costruisce così[9]:

```
crim.mod2 = glm(CRIM ~ ., family=binomial(), data=newdata[,keep.crim])
summary(crim.mod2)
```

La correlazione esistente tra la criminalità e le altre variabili può anche essere spiegata attraverso altri grafici grafici[10].

5 Conclusioni

Fumo

Il modello sulla variabile fumo non si può dire del tutto soddisfacente: in primo luogo, si vede come la variabile di per sé sia poco correlata con le altre variabili del database, in secondo luogo si può verificare dall'R^2 basso del modello definitivo.
Come già evidenziato in precedenza, per quanto riguarda la correlazione diretta con il fumo ha senso parlare di 3 variabili:

- Numero di componenti della famiglia:
  Una delle poche variabili quantitative a tutti gli effetti del database. Quello che emerge dai modelli è che famiglie più numerose tendono ad avere, in termini assoluti, meno fumatori.
  L'interpretazione di questo dato riesce difficile, ma diventa più chiara con l'analisi della prossima variabile.
- Età dell'intervistato:
  La variabile in questione indica, più nello specifico, la fascia d'età cui appartiene l'intervistato. Tuttavia, la distribuzione delle classi non risulta essere molto regolare: le classi che comprendono i primi anni di vita, fino ai 20, hanno 2/3 anni per fascia, mentre poi a salire iniziano ad averne fino a 10 anni per fascia.
  Questa distribuzione particolare delle età potrebbe causare dei problemi di distorsione.
  Quello che emerge è che, più una persona è giovane, meno è portata ad iniziare a fumare. Come conclusione è abbastanza scontata, dal momento che nelle prime fasce gli intervistati non sono ancora legalmente autorizzati a fumare.
  Questa correlazione spiega anche la precedente: infatti, c'è una stretta correlazione tra numero di componenti della famiglia e età degli intervistati: intervistati più giovani tendono a fare parte di famiglie più numerose. E la catena di correlazioni spiega il fenomeno precedente.
- Salute:
  La variabile in questione ha dei valori che non sono di immediata interpretazione dai modelli, a causa di come sono costruiti. La variabile Salute assume valore 1 quando l'intervistato è in salute, e all'aumentare del valore nella variabile corrisponde un peggiore stato di salute dell'intervistato. Pertanto, se la correlazione nel modello risulta positiva, non è perché i fumatori sono più in salute dei non fumatori, ma perché la variabile Salute si muove al "contrario", e di fatto emerge come la salute e il fumo siano correlati negativamente.

[9] L'output del summary nella tabella 10
[10] I grafici sono a fine PDF, figura 1

Criminalità

La variabile criminalità rappresenta la percezione dell'intervistato rispetto al proprio quartiere dal punto di vista della presenza di criminalità, ma non fa affidamento ad indici effettivi calcolati con criteri oggettivi.
Il modello sulla variabile criminalità risulta lievemente più soddisfacente per quanto riguarda la spiegazione della variabile, ma ugualmente insoddisfacente per quanto riguarda l'R^2 del modello. In questo caso, il modello è spiegato da due gruppi di variabili:

- Il rumore, l'inquinamento dell'aria e la sporcizia nel quartiere di residenza dell'intervistato:
  La correlazione con queste variabili è più chiara di quanto sembri, ed è in parte dovuta a come sono state poste le domande: sia la domanda sulla criminalità ("La zona in cui abita presenta rischio di criminali?") e le domande sugli altri argomenti ("La zona in cui abita presenta sporcizia nelle strade / inquinamento dell'aria / rumore?"). Diventa quindi evidente come le risposte siano molto uniformi in base alla percezione dell'intervistato rispetto al proprio quartiere.
  Nel modello la correlazione tra la criminalità e le altre variabili risulta negativa, ma in realtà è strettamente positiva poiché tutte e tre le variabili assumono intensità massima a 1 e decrescente al crescere del valore della variabile.
- La situazione economica della famiglia:
  Ci sono due variabili strettamente correlate con il reddito e sono la variazione della situazione economica (miglioramento / peggioramento) rispetto all'anno precedente e la valutazione delle proprie risorse economiche. Anche in questo caso le correlazioni risultano opposte per la costruzione delle variabili, ma la correlazione che emerge è che famiglie meno agiate tendono a vivere in zone con maggiore criminalità.

Conclusioni generali

In generale non sono stati trovati modelli completamente soddisfacenti, nonostante siano state provate anche altre variabili risposta (come la fiducia: "ti fidi delle persone?") che hanno ritornato approssimativamente lo stesso risultato. Questo potrebbe essere dato dai fini della raccolta dei dati: i dati infatti rappresentano per lo più indici di gradimento e di utilizzo dei servizi, e presentano grandi carenze di risposte, rendendo i dati particolarmente difficili da utilizzare e interpretare. Le domande sono poste in "blocco" per aree tematiche e molto spesso sono anche associate per aree semantiche e costrutti grammaticali che portano a dare risposte simili a domande della stessa area.

6 Tabelle varie

```
[1] "PROFAM"    "PROIND"    "NCOMP"     "ANNO"      "RELPAR"    "ETAMi"     "SESSO"
[8] "STCIVMi"   "ISTRMi"    "TIPNU2"    "NUMNU2"    "RPNUC2"    "TIPFA2Mi"  "REGMF"
[15] "RIPMF"    "COEFIN"   "CITTMi"   "PROSOC"   "GUMED"    "ASSDO"    "RICOV"    "VISMED12"
[23] "ANSANG12" "ACCER12"  "SPOCON"   "CPESO"    "FARM"     "PASTO"    "COLAZ"    "LPRAN"
[31] "PANPAS"   "SALUMI"   "POLLO"    "CBOV"     "CMAIAL"   "LATTE"   "FORM"     "UOVA"
[39] "PESCE"     "VERD"      "FRUTTA"    "LEGUMI"    "PATATE"    "SNACK"     "DOLCI"
[46] "CGRAS"    "FGRAS"    "QTSALE"   "IODIO"    "SALUTE"   "CRONI"    "IPAR"     "RADIO"
[54] "TELE"     "FILMTEL"   "BIBLIO"    "PCTEMPO"   "SENELE"    "GCONT"     "GASBAL"
[61] "GDISPL"   "GCBOL"    "GINF"    "GAS5"     "SCEFO2"   "CALLELGA" "PROINT"   "SPORCO"
[69] "PARCH"    "TRAF"     "INQAR"    "RUMORE"   "CRIM"     "ODSGR"    "ILLSTR"   "CONPAV"
[77] "PARCHI"   "STANZEM"  "TERRAZ"   "GARDEN"   "TELEF"  "RISCAL"   "REACQ1"   "SODACQUA"
[85] "AGFORN"   "AGPRESS"  "AGODOR"   "AGLETTUR" "AGFATTUR" "AGBOLLET" "FOGNA"    "SPEAB"
[93] "ABIPIC"   "ABLONF"   "ACQUA"    "ABICC"    "GODAB"    "FARMA"    "PRSOC"   "UFFPO"
[101] "POLICE"   "UFFCOM"   "MERCAT"   "SMERC"    "CASS"     "CRARIF"   "CERACQ"   "CCARTA"
[109] "CVETRO"   "CFARM"    "CBAT"     "CLATAL"   "CPLAS"    "CRORG"    "CTESSILI" "POAPO"
[117] "R_CARTA"   "R_VETRO"   "R_FARM"    "R_BAT"     "R_LATAL"   "R_PLAS"    "R_RORG"
[124] "R_TESSILI" "ECOSTAZ"  "USOECO2" "SPIDIF1"  "SPIDIF2"  "SPIDIF3"  "SPIDIF4"  "SPIDIF5"
[132]"SPIDIF6"  "SPIDIF7"  "SPIDIF8"  "CAMAB"    "COLFAGG"  "BABYSAGG" "ASSANZAGG" "AVVOC"
[140] "NOTAIO"   "COMMER"   "LAVST"    "LAVATR"   "NLAVAT"  "VIDER"    "VIDEO"    "DVD"
[148] "HIFI"     "SEGTEL"   "FAX"      "TELCOL"   "NTELCO"   "ANTEPA"  "CLIMAT"   "BIC"
[156] "MOTOR"     "AMOTO"     "ABBTV"     "SMARTV"    "AUTO"      "LIBFAM"    "TELCEL"
[163] "NTELCELM"  "TELCIN"    "PC"        "MODEM"     "VGIOC"     "EBOOK"     "MP3"
[170] "FOTODIG"   "NAVSAT"   "AINTERN"   "SITE"      "RISEC"     "FUMO"      "weight"
```

Table 1: Variabili con al più 1750 NA

NCOMP	Numero di componenti della famiglia attuale (variabile quantitativa)
ETAMi	Età in anni compiuti (variabile categorica ma trattata come quantitativa).
SESSO	Sesso (variabile categorica).
REGMF	Regione di residenza (variabile categorica).
CITTMi	Cittadinanza (variabile categorica).
SALUTE	Come va in generale la salute (variabile categorica ma trattata come quantitativa).
SENELE	Soddisfazione sul servizio dell'energia elettrica (variabile categorica ma trattata come quantitativa).
SPORCO	La zona in cui abita presenta sporcizia (variabile categorica ma trattata come quantitativa).
INQAR	La zona in cui abita presenta inquinamento dell'aria (variabile categorica ma trattata come quantitativa).
RUMORE	La zona in cui abita presenta rumore (variabile categorica ma trattata come quantitativa).
CRIM	La zona in cui abita presenta criminalità (variabile categorica ma trattata come quantitativa).
PARCHI	La zona in cui abita presenta parchi a meno di 15 minuti a piedi (variabile categorica).
STANZEM	Di quante stanze si compone l'abitazione (variabile categorica).
GODAB	A che titolo la famiglia occupa l'abitazione (variabile categorica).
AVVOC	La famiglia si è rivolta ad un avvocato negli ultimi 12 mesi (variabile categorica).
BIC	Possesso di bicicletta (variabile categorica).
LIBFAM	Numero di libri in possesso della famiglia (variabile categorica ma trattata come quantitativa).
SITE	Valutazione della situazione economica della famiglia confrontata con quella dell'anno precedente (variabile categorica ma trattata come qualitativa).
RISEC	Come sono state le risorse economiche complessive della famiglia negli ultimi 12 mesi (variabile categorica ma trattata come quantitativa).

Table 2: Variabili mantenute nel database

```
Call:
glm(formula = FUMO ~ ., family = binomial(), data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.728  -1.008  -0.736   1.166   2.043

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.844551   0.164733 -11.197  < 2e-16 ***
NCOMP       -0.036781   0.012876  -2.856 0.004284 **
ETAMi        0.133545   0.006261  21.329  < 2e-16 ***
SESSO2      -0.864158   0.026894 -32.132  < 2e-16 ***
REGMF20      0.001018   0.103318   0.010 0.992136
REGMF30      0.162515   0.064861   2.506 0.012225 *
REGMF40      0.008165   0.073243   0.111 0.911240
REGMF50     -0.137607   0.073499  -1.872 0.061174 .
REGMF60      0.052885   0.084832   0.623 0.533015
REGMF70      0.118556   0.081366   1.457 0.145098
REGMF80      0.114807   0.073207   1.568 0.116826
REGMF90      0.082634   0.073760   1.120 0.262588
REGMF100     0.183668   0.089870   2.044 0.040981 *
REGMF110    -0.014130   0.082392  -0.171 0.863831
REGMF120     0.066084   0.073493   0.899 0.368557
REGMF130    -0.026572   0.083521  -0.318 0.750372
REGMF140    -0.007875   0.094053  -0.084 0.933274
REGMF150    -0.033631   0.071798  -0.468 0.639494
REGMF160    -0.182245   0.077253  -2.359 0.018320 *
REGMF170     0.013008   0.089657   0.145 0.884646
REGMF180    -0.228266   0.085883  -2.658 0.007863 **
REGMF190    -0.039002   0.077592  -0.503 0.615209
REGMF200     0.061486   0.083398   0.737 0.460961
CITTMi3     -0.193348   0.080942  -2.389 0.016907 *
SALUTE       0.027068   0.019366   1.398 0.162191
SENELE       0.081349   0.022911   3.551 0.000384 ***
SPORCO      -0.048955   0.018020  -2.717 0.006594 **
INQAR       -0.009223   0.021767  -0.424 0.671780
RUMORE       0.018703   0.021186   0.883 0.377349
CRIM         0.011462   0.019829   0.578 0.563233
PARCHI2      0.115641   0.035277   3.278 0.001045 **
STANZEM     -0.009230   0.009431  -0.979 0.327721
GODAB2      -0.408090   0.042795  -9.536  < 2e-16 ***
GODAB3      -0.239142   0.105059  -2.276 0.022831 *
GODAB4      -0.188919   0.071923  -2.627 0.008622 **
GODAB5      -0.296758   0.145238  -2.043 0.041027 *
AVVOC        0.222898   0.038847   5.738 9.59e-09 ***
BIC8        -0.032186   0.031377  -1.026 0.304994
LIBFAM       0.040045   0.007338   5.457 4.84e-08 ***
SITE        -0.003245   0.022065  -0.147 0.883092
RISEC        0.125527   0.028779   4.362 1.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34501  on 25436  degrees of freedom
Residual deviance: 32403  on 25396  degrees of freedom
AIC: 32485

Number of Fisher Scoring iterations: 4
```

Table 3: Summary modello con tutte le variabili

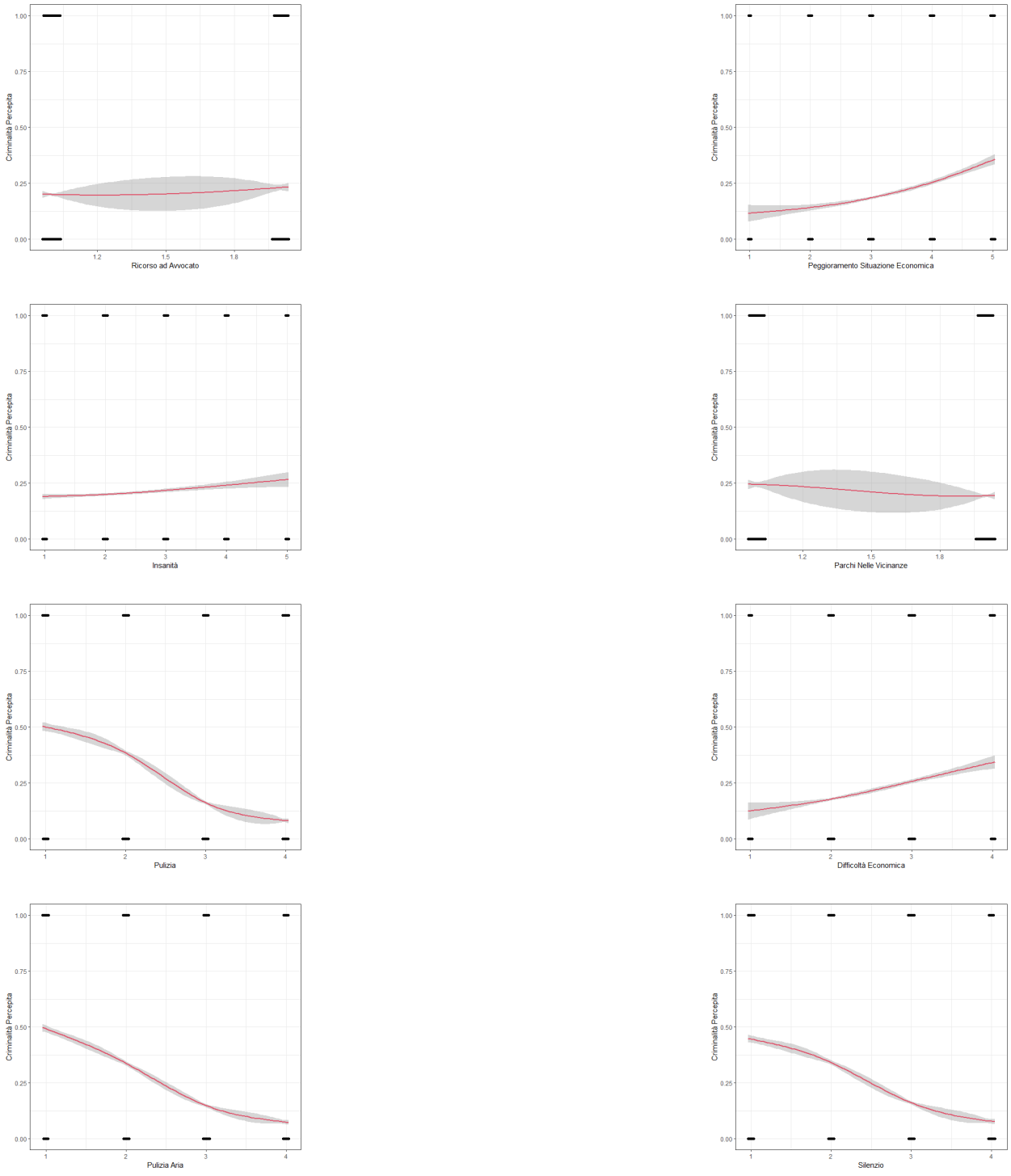


Figure 1: Grafici di correlazione con la criminalità

```
Step:  AIC=32476.13
FUMO ~ NCOMP + ETAMi + SESSO + REGMF + CITTMi + SENELE + SPORCO +
    PARCHI + GODAB + AVVOC + LIBFAM + RISEC

          Df Deviance   AIC
<none>          32408 32476
+ SALUTE   1    32406 32476
+ BIC      1    32407 32477
+ STANZEM  1    32407 32477
+ RUMORE   1    32407 32477
+ CRIM     1    32408 32478
+ INQAR    1    32408 32478
+ SITE     1    32408 32478
- CITTMi   1    32414 32480
- SPORCO   1    32415 32481
- PARCHI   1    32419 32485
- NCOMP    1    32420 32486
- SENELE   1    32421 32487
- REGMF   19    32466 32496
- RISEC    1    32434 32500
- LIBFAM   1    32435 32501
- AVVOC    1    32440 32506
- GODAB    4    32513 32573
- ETAMi    1    33036 33102
- SESSO    1    33468 33534
```

Table 4: Ultimo passo dello step AIC

```
Call:
glm(formula = FUMO ~ NCOMP + ETAMi + SESSO + REGMF + CITTMi +
    SENELE + SPORCO + PARCHI + GODAB + AVVOC + LIBFAM + RISEC,
    family = binomial(), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7135  -1.0085  -0.7371   1.1672   2.0515

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.8216495  0.1501270 -12.134  < 2e-16 ***
NCOMP       -0.0419110  0.0123634  -3.390 0.000699 ***
ETAMi        0.1374471  0.0056147  24.480  < 2e-16 ***
SESSO2      -0.8601698  0.0267756 -32.125  < 2e-16 ***
REGMF20      0.0036918  0.1031591   0.036 0.971452
REGMF30      0.1611709  0.0646459   2.493 0.012662 *
REGMF40     -0.0002904  0.0729683  -0.004 0.996825
REGMF50     -0.1502382  0.0730122  -2.058 0.039617 *
REGMF60      0.0427088  0.0842816   0.507 0.612337
REGMF70      0.1232133  0.0807961   1.525 0.127261
REGMF80      0.1023429  0.0728310   1.405 0.159958
REGMF90      0.0771526  0.0734625   1.050 0.293612
REGMF100     0.1811413  0.0895602   2.023 0.043118 *
REGMF110    -0.0187833  0.0819812  -0.229 0.818778
REGMF120     0.0750532  0.0728657   1.030 0.303000
REGMF130    -0.0297329  0.0828824  -0.359 0.719792
REGMF140    -0.0012250  0.0930062  -0.013 0.989491
REGMF150    -0.0282783  0.0709887  -0.398 0.690373
REGMF160    -0.1852308  0.0771032  -2.402 0.016289 *
REGMF170     0.0253566  0.0887942   0.286 0.775210
REGMF180    -0.2168809  0.0844939  -2.567 0.010263 *
REGMF190    -0.0307877  0.0768538  -0.401 0.688714
REGMF200     0.0723176  0.0820443   0.881 0.378077
CITTMi3     -0.1942964  0.0808072  -2.404 0.016197 *
SENELE       0.0804811  0.0228455   3.523 0.000427 ***
SPORCO      -0.0430932  0.0161435  -2.669 0.007599 **
PARCHI2      0.1140211  0.0352189   3.238 0.001206 **
GODAB2      -0.4177329  0.0421111  -9.920  < 2e-16 ***
GODAB3      -0.2457531  0.1047684  -2.346 0.018992 *
GODAB4      -0.1939670  0.0717222  -2.704 0.006842 **
GODAB5      -0.3030588  0.1450484  -2.089 0.036675 *
AVVOC        0.2197698  0.0387523   5.671 1.42e-08 ***
LIBFAM       0.0370942  0.0071392   5.196 2.04e-07 ***
RISEC        0.1283659  0.0253958   5.055 4.31e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34501  on 25436  degrees of freedom
Residual deviance: 32408  on 25403  degrees of freedom
AIC: 32476

Number of Fisher Scoring iterations: 4
```

Table 5: Summary del modello dopo lo step AIC

```
Call:
glm(formula = FUMO ~ ETAMi + AVVOC + SESSO + RISEC + SENELE +
    SPORCO + LIBFAM, family = binomial(), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6603  -1.0161  -0.7594   1.1787   2.0099

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.287634   0.122476 -18.678  < 2e-16 ***
ETAMi        0.141242   0.004980  28.361  < 2e-16 ***
AVVOC        0.214549   0.038511   5.571 2.53e-08 ***
SESSO2      -0.848965   0.026587 -31.932  < 2e-16 ***
RISEC        0.156578   0.024890   6.291 3.16e-10 ***
SENELE       0.069968   0.022541   3.104  0.00191 **
SPORCO      -0.036843   0.015308  -2.407  0.01610 *
LIBFAM       0.034751   0.006765   5.137 2.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34501  on 25436  degrees of freedom
Residual deviance: 32621  on 25429  degrees of freedom
AIC: 32637

Number of Fisher Scoring iterations: 4
```

Table 6: Summary del modello con le variabili significative

```
Call:
glm(formula = CRIM ~ ., family = binomial(), data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0356  -0.6459  -0.4246  -0.2292   2.9781

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.296579   0.201840   6.424 1.33e-10 ***
NCOMP        0.016943   0.016760   1.011 0.312046
ETAMi        0.004649   0.007977   0.583 0.560085
SESSO2       0.068969   0.035647   1.935 0.053018 .
REGMF20     -0.494003   0.179652  -2.750 0.005964 **
REGMF30      0.451399   0.081793   5.519 3.41e-08 ***
REGMF40     -0.415403   0.113310  -3.666 0.000246 ***
REGMF50      0.304762   0.095297   3.198 0.001384 **
REGMF60     -0.523415   0.142041  -3.685 0.000229 ***
REGMF70     -0.122211   0.106693  -1.145 0.252025
REGMF80      0.605377   0.090584   6.683 2.34e-11 ***
REGMF90      0.127316   0.097807   1.302 0.193016
REGMF100    -0.096980   0.123902  -0.783 0.433792
REGMF110     0.294625   0.107501   2.741 0.006131 **
REGMF120     0.358764   0.089265   4.019 5.84e-05 ***
REGMF130     0.116511   0.111403   1.046 0.295628
REGMF140    -0.881233   0.167071  -5.275 1.33e-07 ***
REGMF150     0.548030   0.086488   6.337 2.35e-10 ***
REGMF160     0.125911   0.094932   1.326 0.184729
REGMF170    -0.436277   0.132746  -3.287 0.001014 **
REGMF180    -0.163381   0.113515  -1.439 0.150068
REGMF190    -0.193444   0.098765  -1.959 0.050157 .
REGMF200    -0.854819   0.137795  -6.204 5.52e-10 ***
CITTMi3     -0.357075   0.105090  -3.398 0.000679 ***
SALUTE       0.044616   0.025193   1.771 0.076567 .
SENELE       0.019551   0.029777   0.657 0.511449
SPORCO      -0.655999   0.022027 -29.781  < 2e-16 ***
INQAR       -0.454231   0.026735 -16.990  < 2e-16 ***
RUMORE      -0.265271   0.026219 -10.117  < 2e-16 ***
PARCHI2     -0.183652   0.044691  -4.109 3.97e-05 ***
STANZEM     -0.012754   0.012878  -0.990 0.321994
GODAB2      -0.217802   0.053466  -4.074 4.63e-05 ***
GODAB3      -0.500826   0.143146  -3.499 0.000468 ***
GODAB4      -0.237402   0.095802  -2.478 0.013210 *
GODAB5       0.090850   0.175763   0.517 0.605235
AVVOC        0.152636   0.049129   3.107 0.001891 **
BIC8        -0.103570   0.040773  -2.540 0.011080 *
LIBFAM      -0.022559   0.009612  -2.347 0.018927 *
SITE         0.252803   0.028481   8.876  < 2e-16 ***
RISEC        0.113189   0.036352   3.114 0.001848 **
FUMO         0.023812   0.036543   0.652 0.514655
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25745  on 25436  degrees of freedom
Residual deviance: 21110  on 25396  degrees of freedom
AIC: 21192

Number of Fisher Scoring iterations: 5
```

Table 7: Summary del primo modello sulla criminalità

```
Step:  AIC=21185.06
CRIM ~ SESSO + REGMF + CITTMi + SALUTE + SPORCO + INQAR + RUMORE +
    PARCHI + GODAB + AVVOC + BIC + LIBFAM + SITE + RISEC

          Df Deviance   AIC
<none>          21113 21185
- SESSO    1    21116 21186
+ STANZEM  1    21112 21186
+ FUMO     1    21113 21187
+ NCOMP    1    21113 21187
+ SENELE   1    21113 21187
+ ETAMi    1    21113 21187
- SALUTE   1    21118 21188
- LIBFAM   1    21119 21189
- BIC      1    21120 21190
- AVVOC    1    21123 21193
- RISEC    1    21124 21194
- CITTMi   1    21125 21195
- PARCHI   1    21130 21200
- GODAB    4    21140 21204
- SITE     1    21194 21264
- RUMORE   1    21215 21285
- INQAR    1    21408 21478
- REGMF   19    21572 21606
- SPORCO   1    22037 22107
```

Table 8: Ultimo step dello step AIC sulla criminalità

```
Call:
glm(formula = CRIM ~ SESSO + REGMF + CITTMi + SALUTE + SPORCO +
    INQAR + RUMORE + PARCHI + GODAB + AVVOC + BIC + LIBFAM +
    SITE + RISEC, family = binomial(), data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0244  -0.6452  -0.4250  -0.2291   2.9616

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.378209   0.173025   7.965 1.65e-15 ***
SESSO2       0.062795   0.034885   1.800 0.071850 .
REGMF20     -0.499626   0.179615  -2.782 0.005408 **
REGMF30      0.452170   0.081748   5.531 3.18e-08 ***
REGMF40     -0.419930   0.113090  -3.713 0.000205 ***
REGMF50      0.296981   0.094867   3.130 0.001745 **
REGMF60     -0.531947   0.141760  -3.752 0.000175 ***
REGMF70     -0.123580   0.106634  -1.159 0.246491
REGMF80      0.600996   0.090390   6.649 2.95e-11 ***
REGMF90      0.122250   0.097459   1.254 0.209708
REGMF100    -0.099234   0.123514  -0.803 0.421732
REGMF110     0.292916   0.106954   2.739 0.006168 **
REGMF120     0.363111   0.089049   4.078 4.55e-05 ***
REGMF130     0.113261   0.110978   1.021 0.307458
REGMF140    -0.884763   0.166670  -5.308 1.11e-07 ***
REGMF150     0.557642   0.085433   6.527 6.70e-11 ***
REGMF160     0.129670   0.094677   1.370 0.170812
REGMF170    -0.431848   0.132498  -3.259 0.001117 **
REGMF180    -0.160570   0.112973  -1.421 0.155225
REGMF190    -0.189015   0.098215  -1.925 0.054292 .
REGMF200    -0.852795   0.137321  -6.210 5.29e-10 ***
CITTMi3     -0.352873   0.104977  -3.361 0.000775 ***
SALUTE       0.048922   0.022177   2.206 0.027383 *
SPORCO      -0.657314   0.021965 -29.926  < 2e-16 ***
INQAR       -0.455962   0.026689 -17.084  < 2e-16 ***
RUMORE      -0.265251   0.026209 -10.121  < 2e-16 ***
PARCHI2     -0.182676   0.044668  -4.090 4.32e-05 ***
GODAB2      -0.223593   0.052399  -4.267 1.98e-05 ***
GODAB3      -0.504975   0.142814  -3.536 0.000406 ***
GODAB4      -0.242890   0.095616  -2.540 0.011077 *
GODAB5       0.084367   0.175556   0.481 0.630822
AVVOC        0.151931   0.048959   3.103 0.001914 **
BIC8        -0.100602   0.039128  -2.571 0.010138 *
LIBFAM      -0.023587   0.009438  -2.499 0.012446 *
SITE         0.254159   0.028430   8.940  < 2e-16 ***
RISEC        0.120393   0.035690   3.373 0.000743 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25745  on 25436  degrees of freedom
Residual deviance: 21113  on 25401  degrees of freedom
AIC: 21185

Number of Fisher Scoring iterations: 5
```

Table 9: Summary del modello dopo lo step AIC

```
Call:
glm(formula = CRIM ~ ., family = binomial(), data = newdata[,
    keep.crim])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9736  -0.6439  -0.4531  -0.2641   2.7485

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.49313    0.14818  10.077  < 2e-16 ***
CITTMi3     -0.22209    0.10331  -2.150 0.031577 *
SALUTE       0.05595    0.02155   2.597 0.009413 **
SPORCO      -0.63928    0.02070 -30.881  < 2e-16 ***
INQAR       -0.53814    0.02544 -21.153  < 2e-16 ***
RUMORE      -0.24422    0.02561  -9.537  < 2e-16 ***
PARCHI2     -0.18306    0.04107  -4.457 8.30e-06 ***
GODAB2      -0.26069    0.05078  -5.134 2.84e-07 ***
GODAB3      -0.43915    0.14019  -3.133 0.001732 **
GODAB4      -0.32394    0.09373  -3.456 0.000548 ***
GODAB5       0.05730    0.17420   0.329 0.742226
SITE         0.24330    0.02800   8.688  < 2e-16 ***
RISEC        0.14183    0.03497   4.056 5.00e-05 ***
AVVOC        0.11718    0.04801   2.441 0.014655 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25745  on 25436  degrees of freedom
Residual deviance: 21590  on 25423  degrees of freedom
AIC: 21618

Number of Fisher Scoring iterations: 5
```

Table 10: Summary del modello definitivo