Introduzione a R
Esempio di regressione multipla: tasso di criminalità negli Stati Uniti

Questa nota contiene una sequenza di istruzioni in R, le istruzioni sono stampate in carattere diverso e vanno ricopiate esttamente sulla riga di comando.
Il dataset contiene alcune statistiche demografiche e sociali per 47 stati americani nel 1960. La variabile che si vorrebbe spiegata dalle eltre è il tasso di criminalità (R).

R : tasso di criminalità, # di crimini denunciati pe rmilione di abitanti

Age : maschi di età 14-24 per 1000 abitanti

S : indicatore per "Stato del Sud" (0 = No, 1 = Yes)

Ed : 10×# medio di anni di scuola per persone di età uguale o superiore a 25 anni

Ex0 : 1960 spesa procapite per la polizia (stato e governo locale)

Ex1 : 1959 spesa procapite per la polizia (stato e governo locale)

LF : partecipazione alla forza lavoro per 1000 maschi di età 14-24

M : numero di maschi per 1000 femmine

N : popolazione dello stato (× 100000)

NW : numero di non bianchi per 1000

U1 : disoccupati per 1000 maschi di età 14-24 nelle città

U2 : disoccupati per 1000 maschi di età 35-39 nelle città

W : valore mediano del patrimonio familiare (×10 $)

X : numero di famiglie su 1000 che guadagnano meno di metà del reddito mediano (dello stato)

A - Caricare il file

1. `usc=read.table("C:/.../UScrime.txt",header=T)`
2. `names(usc)`

3. `usc[1:4,]`

B - Guardiamo i dati

1. raffigurazioni grafiche: tutte quantitative tranne `S`
2. `pairs(usc)`
3. `cor(usc)`

C - Il problema della collinearità

1. `plot(usc$Ex0,usc$R)`
2. `plot(usc$Ex1,usc$R)` Ex0 e Ex1 sono positivamente correlate con R.
3. `summary(lm(R~Ex0,data=usc))`
4. `summary(lm(R~Ex1,data=usc))`
5. `summary(lm(R~Ex1+Ex0,data=usc))` confrontare il risultato con i due precedenti, cosa è successo? Notare che il coefficiente di Ex0 ha cambiato segno
6. `plot(usc$Ex1,usc$Ex0)` ecco il problema!
7. `summary(lm(Ex0~Ex1,data=usc))` la (confermata) collinearità produce i risultati delle due precedenti regressioni
8. anche le coppie `W`,`X` e `U1`,`U2` presentano lo stesso problema.

D - Regressione multipla

1. `summary(usc.fit1=lm(R~1,data=usc))`
2.
```
step(usc.fit1,
     scope=list(lower=as.formula("R~1"),
                upper=as.formula("R~Age+S+Ed+Ex0+Ex1+LF+M+N+NW+U1+U2+W+X")))
```
3. `summary(usc.fit1=lm(R~Ex0+X+Age+W+Ed+U2,data=usc))` ha senso?
4. `summary(usc.fit1=lm(R~Ex0+X+Age+Ed+U2,data=usc))` ha senso?
5. `par(mfrow=c(2,2))` 4 grafici in una finestra
6. `plot(usc.fit1)`
7. ci sono vari problemi di interpretazione:
   - la variabile `Ed` ha un effetto positivo sul tasso di criminalità
   - `Ex0` e `R`, qual è il nesso di causalità?