

```
#####  
# Model selection: metodi Backward e Forward  
#####  
data = read.table(file.choose(), header = TRUE) # covariates_useless2.txt  
head(data)  
str(data)  
  
# separo le variabili  
spesa = data$spesa # spesa alimentare  
reddito = data$reddito # reddito  
figli = data$figli # numero figli  
metri = data$metri # metri quadri abitazione  
sesso = data$sesso # sesso del capofamiglia (1=M)  
genitori = data$genitori # dummy per entrambi i genitori che lavorano (1=sì)  
zona = data$zona # zona di abitazione (1=Urbana)  
  
# PROCEDURA BACKWARD:  
  
# Stimo il modello completo (con tutte le covariate)  
mod = lm(spesa ~ reddito + figli + metri + sesso + genitori + zona)  
summary(mod)  
  
# Siccome la variabile con il p-value più elevato (e > 0.05) è sesso, elimino sesso e stimo il modello  
senza sesso.  
mod1 = lm(spesa ~ reddito + figli + metri + genitori + zona)  
summary(mod1)  
  
# Adesso, il p-value più grande (e maggiore di 0.05) è quello della covariata metri. Quindi, elimino  
metri.  
mod2 = lm(spesa ~ reddito + figli + genitori + zona)  
summary(mod2)  
  
# A questo punto tutte le covariate inserite in mod2 risultano statisticamente significative al 5%.  
Pertanto, non ci sono altri candidati all'eliminazione. Il modello definitivo è mod2:  
  
# spesa ~ reddito + zona+ figli +genitori  
  
# Qualora il livello di significatività sia fissato all'1%, anche la variabile genitori andrebbe  
eliminata.  
#####  
# PROCEDURA FORWARD:  
  
# Stimo un modello per ciascuna covariata candidata alla selezione  
mod1 = lm(spesa ~ reddito)  
mod2 = lm(spesa ~ figli)  
mod3 = lm(spesa ~ metri)  
mod4 = lm(spesa ~ sesso)  
mod5 = lm(spesa ~ genitori)  
mod6 = lm(spesa ~ zona)  
summary(mod1)  
summary(mod2)  
summary(mod3)  
summary(mod4)  
summary(mod5)  
summary(mod6)  
  
# Sia reddito che zona presentano i p-values più piccoli, inoltre reddito presenta un R-quadro  
aggiustato più elevato. Quindi aggiungo reddito.  
  
# Stimo un modello (con la variabile reddito) per ciascuna altra variabile candidata alla selezione  
mod11 = lm(spesa ~ reddito + figli)  
mod21 = lm(spesa ~ reddito + metri)  
mod31 = lm(spesa ~ reddito + sesso)  
mod41 = lm(spesa ~ reddito + genitori)  
mod51 = lm(spesa ~ reddito + zona)  
summary(mod11)  
summary(mod21)  
summary(mod31)
```

```
summary(mod41)
summary(mod51)

# La variabile zona presenta il p-value più piccolo e l'R-quadro aggiustato più elevato. Quindi
aggiungo zona.
mod12 = lm(spesa ~ reddito + zona+figli)
mod22 = lm(spesa ~ reddito + zona+metri)
mod32 = lm(spesa ~ reddito + zona+sexso)
mod42 = lm(spesa ~ reddito + zona+genitori)
summary(mod12)
summary(mod22)
summary(mod32)
summary(mod42)

# La variabile figli presenta il p-value più piccolo e l'R-quadro aggiustato più elevato. Quindi
aggiungo figli.
mod13 = lm(spesa ~ reddito + zona+ figli +metri)
mod23 = lm(spesa ~ reddito + zona+ figli +sexso)
mod33 = lm(spesa ~ reddito + zona+ figli +genitori)
summary(mod13)
summary(mod23)
summary(mod33)

# La variabile genitori è significativa al 5%. Inserisco genitori.
mod14 = lm(spesa ~ reddito + zona+ figli +genitori+metri)
mod24 = lm(spesa ~ reddito + zona+ figli +genitori+sexso)
summary(mod14)
summary(mod24)

# Nessuna altra variabile è significativa. Quindi il modello selezionato è:
# spesa ~ reddito + zona+ figli +genitori

# NOTA BENE: qualora tra le variabili esplicative ci siano uno o più fattori con più di due livelli,
l'eliminazione (procedura backward) o l'inserimento (procedura forward) di ciascun fattore nel modello
deve essere valutato per tutti i livelli contemporaneamente: è, pertanto, necessario ricorrere ad un
test F che verifichi l'ipotesi nulla che tutti i coefficienti di regressione relativi alle dummies
dello stesso fattore sono uguali a 0

#####

# Diagnosi della Multicollinearità

#####

### carico i dati
data = read.table(file.choose(),header=T) # multicollinear.txt
str(data)
x1 = data$x1
x2 = data$x2
x3 = data$x3
y = data$y

# Calcolo i coefficienti di correlazione tra le coppie di variabili esplicative
cor(x2, x3)
cor(x1, x3)
cor(x1, x2)
# Indizio num. 1: La correlazione tra x2 e x3 è molto elevata: sospetto di multicollinearità

# Stimo il modello di regressione multipla
mod = lm(y ~ x1+x2+x3)
summary(mod)

# Indizio num. 2: Noto che l'R-quadro aggiustato del modello è molto elevato e, contemporaneamente,
due coefficienti di regressione su tre non sono significativamente diversi da zero: è probabile che
sia di fronte ad una situazione di multicollinearità.

# Stimo il modello con due sole covariate
mod = lm(y ~ x1+x2)
summary(mod)

# Indizio num. 3: Togliendo x3 noto che x2 diventa significativa: altro indizio di multicollinearità
```

```
# Regredisco x1 su x2 e x3
mod = lm(x1 ~ x2+x3)
summary(mod)

# Calcolo VIF1
re1 = mod$residuals
r21 = 1-sum(re1^2)/var(x1)
VIF1 = 1/(1-r21)
VIF1

# Regredisco x2 su x1 e x3
mod = lm(x2 ~ x1+x3)
summary(mod)

# Calcolo VIF2
re2 = mod$residuals
r22 = 1-sum(re2^2)/var(x2)
VIF2 = 1/(1-r22)
VIF2

# Regredisco x3 su x1 e x2
mod = lm(x3 ~ x2+x1)
summary(mod)

# Calcolo VIF3
re3 = mod$residuals
r23 = 1-sum(re3^2)/var(x3)
VIF3 = 1/(1-r23)
VIF3

# VIF medio
VIFm = (VIF1+VIF2+VIF3)/3
VIFm

# Essendo il VIF medio più grande di 1, posso concludere che sussiste multicollinearità, causata dalla
compresenza nel modello delle variabili x2 e x3.

#####

# Regressione Multivariata

#####

# carico i dati
data = read.table(file.choose(),header=T,sep="\t") # regr_multivariata.txt
str(data)

reddito = data$reddito
anni = data$anni_lavoro
eta = data$eta
sesso = data$sesso
titolo = data$titolo
summary(titolo)

# stimo il modello per il reddito
mod1 = lm(reddito ~ eta+sesso+titolo)
summary(mod1)

# stimo il modello per gli anni di lavoro
mod2 = lm(anni ~ eta+sesso+titolo)
summary(mod2)
anova(mod2) # verifico se titolo è significativa nel suo complesso

# matrice disegno (è uguale per entrambi i modelli perché utilizzo le stesse covariate)
X = model.matrix(mod1)
# oppure: X = model.matrix(mod2)
head(X)
```

```

# creo vettore delle variabili risposta
Y = cbind(reddito,anni)
head(Y)

# Calcolo i coefficienti di regressione
Bh = solve(t(X)%*%X)%*%t(X)%*%Y
Bh

# Calcolo i valori previsti delle due variabili risposta
Yh = X%*%Bh
head(Yh)

# Residui
Eh = Y-Yh
head(Eh)

# Calcolo la matrice di varianze e covarianze
n = nrow(Yh)
npar = dim(X)[2] # è il numero di parametri stimati per ogni Y (= num. covariate + intercetta)
Sih = t(Eh)%*%Eh/(n-npar)
Sih

# Correlazione tra reddito ed anni di lavoro
Cov = Sih[1,2]
Cov
Si1 = sqrt(Sih[1,1])
Si1
Si2 = sqrt(Sih[2,2])
Si2
Corr = Cov/(Si1*Si2)
Corr

#####

# Regressione logistica (o logit)

#####

data = read.table(file.choose(),header=T,sep="\t")# binary.txt

str(data)
admit = data$admit
gre = data$gre
gpa = data$gpa
rank = data$rank

# admit = factor(admit)

# creo le dummies per il prestigio dell'università
zrank = as.factor(rank)

# Stimo il modello logit
?glm
mod_logit = glm(formula = admit ~ gre + gpa + zrank, family = binomial(link="logit"))
summary(mod_logit)

# Ricorda che i coefficienti di regressione nel modello logistico indicano la variazione attesa del
log-odds della variabile risposta a fronte di un incremento unitario della variabile esplicativa
corrispondente.
# Esempio 1: a fronte di un incremento di un punto sul gpa, il log-odds di essere ammessi alla
graduate school aumenta di 0.804.
# Esempio 2: l'aver frequentato una undergraduate school di rango 2 rispetto ad una di rango 1 fa sì
che il log-odds di essere ammessi alla graduate school si riduca di 0.675.

# Stimo gli intervalli di confidenza dei coefficienti di regressione
confint(mod_logit)

# Per facilitare l'interpretazione del modello è utile calcolare l'esponenziale dei coefficienti di
regressione, così da interpretarli come odds-ratios:
exp_beta = exp(mod_logit$coefficients)
exp_beta

```

```
# Esempio 1: a fronte di un incremento di 1 punto sul gpa, l'odds di essere ammessi alla graduate
school aumenta di 2.235.
# Esempio 2: l'aver frequentato una undergraduate school di rango 2 rispetto ad una di rango 1 fa sì
che l'odds di essere ammessi alla graduate school è pari a 0.51, cioè la probabilità di essere ammessi
rispetto alla probabilità di non essere ammessi per uno studente che ha frequentato una scuola di
rango 2 è circa la metà (0.51) della probabilità di essere ammessi rispetto alla probabilità di non
essere ammessi per uno studente che ha frequentato una scuola di rango 1.

# Stimo gli intervalli di confidenza per exp_beta
exp(confint(mod_logit))
# Osservo che l'effetto di gre è quasi nullo (benché statisticamente significativo al 5%), essendo il
coefficiente beta prossimo a zero (e il corrispondente esponenziale prossimo ad 1). Pertanto, valuto
l'ipotesi di eliminare tale variabile dal modello, tramite il test della devianza.
# Stimo il modello senza gre
mod_logit2 = glm (formula = admit ~ gpa + zrank, family = binomial (link="logit"))
summary(mod_logit2)

# Test della devianza:
anova(mod_logit2, mod_logit, test = "Chisq")

# Accetto l'ipotesi nulla di differenza non significativa (al 5%) tra i due modelli a confronto
(quello con gre e quello senza). Pertanto preferisco il modello più parsimonioso, cioè quello senza la
variabile gre.

# Calcolo le stime delle probabilità individuali di essere ammessi ad una graduate school
prob = predict(mod_logit2, type= "response")
prob

# Stimo il modello probit
mod_probit = glm (formula = admit ~ gre + gpa + zrank, family = binomial (link="probit"))
summary(mod_probit)
```