1. **ALLINEAMENTO DI FATTORI, ESEMPI SINTETICI**

**Esercizio**. Creare una tabella con dati sintetici in cui sia percepibile il fenomeno dell’allineamento ed esaminare le sue conseguenze.

Consideriamo il modello

 Y = X1 + X2 + 0.5\*X3 – 0.5\*X4 + sig1\*eps1

dove X1,X3,X4 sono normali standard indipendenti e

 X2 = X1 + sig2\*eps2

e dove, infine, eps1 ed eps2 sono normali standard indipendenti.

Nrow=20; Ncol=5;sig1=0.5 ;sig2=0.1

A = matrix(nrow=Nrow , ncol=Ncol )

A[,1] = rnorm(Nrow)

A[,2] = A[,1] + sig2\*rnorm(Nrow)

A[,3] = rnorm(Nrow)

A[,4] = rnorm(Nrow)

A[,5]= A[,1]+ A[,2]+ 0.5\*A[,3]-0.5\* A[,4]+ sig1\*rnorm(Nrow)

Reg.prova = lm(A[,5]~ A[,1]+ A[,2]+ A[,3]+ A[,4])

summary(Reg.prova)

Siete invitati a cambiare un po’ di parametri dell’esercizio (es. aumentare Nrow; diminuire sig2; diminuire i coefficienti di X3 ed X4).

Eliminamo X2:

Nrow=20; Ncol=5;sig1=0.5 ;sig2=0.1

A = matrix(nrow=Nrow , ncol=Ncol )

A[,1] = rnorm(Nrow)

A[,2] = A[,1] + sig2\*rnorm(Nrow)

A[,3] = rnorm(Nrow)

A[,4] = rnorm(Nrow)

A[,5]= A[,1]+ A[,2]+ 0.5\*A[,3]-0.5\* A[,4]+ sig1\*rnorm(Nrow)

Reg.prova = lm(A[,5]~ A[,1]+ A[,2]+ A[,3]+ A[,4])

summary(Reg.prova)

Reg.prova2 = lm(A[,5]~ A[,1]+ A[,3]+ A[,4])

summary(Reg.prova2)

Si confrontino i risultati.

1. **PROSEGUIMENTO DEL METODO DI ELIMINAZIONE DI FATTORI (RLM)**

A <- read.table ('clipboard', header=TRUE)

 PLIC SC SA.SC TD TMI

Piem 0.088 0.471 -0.707 -0.607 -0.3950

Vaos -1.545 0.348 -0.642 -0.813 1.5780

Lomb 0.202 1.397 -0.836 -0.790 -0.5380

TrAA 0.677 0.435 -1.269 -0.966 -0.0750

Vene 0.088 1.334 -1.210 -0.848 -0.4970

FrVG 0.639 -0.005 -1.028 -0.804 -1.3010

Ligu 1.190 -0.247 0.470 -0.429 -0.3540

EmRo 0.658 1.177 -1.315 -0.863 -0.3470

Tosc 0.126 1.092 -0.795 -0.644 -1.3550

Umbr -1.431 0.675 -0.140 -0.524 -1.2870

Marc 0.278 1.090 -0.265 -0.702 -0.0006

Lazi 2.329 0.546 -0.080 -0.113 -0.0140

Abru 0.335 -0.373 0.402 -0.456 0.0400

Moli 0.658 -1.289 0.065 0.451 -1.1510

Camp -1.811 -1.314 2.031 1.664 0.4140

Pugl -0.766 -0.926 1.038 0.648 1.1090

Basi -0.747 -1.154 0.661 0.844 2.0010

Cala -0.500 -1.727 1.571 2.153 0.6320

Sici -0.918 -1.130 1.332 1.517 1.7830

Sard 0.449 -0.403 0.717 1.285 -0.2380

cor(A)

 PLIC SC SA.SC TD TMI

PLIC 1.0000000 0.3223197 -0.4110268 -0.3664348 -0.4432647

SC 0.3223197 1.0000000 -0.8417383 -0.8501667 -0.4834029

SA.SC -0.4110268 -0.8417383 1.0000000 0.9054169 0.5136762

TD -0.3664348 -0.8501667 0.9054169 1.0000000 0.4868433

TMI -0.4432647 -0.4834029 0.5136762 0.4868433 1.0000000

PLIC = A[,1]

SC = A[,2]

SA.SC = A[,3]

TD = A[,4]

TMI = A[,5]

------------------------ richiamo ------------------

Reg2 = lm(TD~SC+SA.SC+TMI)

> summary(Reg2)

Call:

lm(formula = TD ~ SC + SC.SA + TMI)

Residuals:

 Min 1Q Median 3Q Max

-0.8290 -0.2070 0.0137 0.2329 0.7014

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.000103 0.095430 0.001 0.9992

SC -0.300484 0.182547 -1.646 0.1192

SC.SA 0.647871 0.186253 3.478 0.0031 \*\*

TMI 0.008717 0.114857 0.076 0.9404

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4268 on 16 degrees of freedom

Multiple R-squared: 0.8464, Adjusted R-squared: 0.8176

F-statistic: 29.4 on 3 and 16 DF, p-value: 9.597e-07

--------------------- fine richiamo ----------------

Reg3 = lm(TD~SC+SA.SC)

summary(Reg3)

Call:

lm(formula = TD ~ SC + SC.SA)

Residuals:

 Min 1Q Median 3Q Max

-0.83048 -0.20756 0.00737 0.22957 0.69637

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.0001047 0.0925974 0.001 0.99911

SC -0.3020096 0.1760508 -1.715 0.10443

SC.SA 0.6510650 0.1760519 3.698 0.00179 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4141 on 17 degrees of freedom

Multiple R-squared: 0.8464, Adjusted R-squared: 0.8283

F-statistic: 46.83 on 2 and 17 DF, p-value: 1.216e-07

>

R^2 non è diminuito. R^2 adjusted è migliorato ancora. Persino i singoli p-values sono migliorati.

Togliamo anche l’intercetta:

Reg.int = lm(TD~SC+SA.SC+0)

summary(Reg.int)

E’ migliorato ancora. Nel seguito verrà lasciata.

A questo punto è naturale fermarsi, per varie ragioni: due fattori non sono certo troppi, entrambi hanno p-value accettabili, entrambi hanno correlazione accettabile con TD.

Però, a titolo di studio, eliminiamo il peggiore, SC:

Reg4 = lm(TD~SA.SC)

summary(Reg4)

> summary(Reg4)

Call:

lm(formula = TD ~ SC.SA)

Residuals:

 Min 1Q Median 3Q Max

-0.85463 -0.24693 0.05422 0.26318 0.73066

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.00015 0.09747 0.002 0.999

SC.SA 0.90528 0.10005 9.049 4.06e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4359 on 18 degrees of freedom

Multiple R-squared: 0.8198, Adjusted R-squared: 0.8098

F-statistic: 81.88 on 1 and 18 DF, p-value: 4.064e-08

[Si provi anche togliere il migliore: Reg5 = lm(TD~SC); summary(Reg5) ]

>

Il p-value del fattore rimasto è molto migliorato: coincide ora col p-value globale. Ma R^2 è diminuito ed anche R^2 adjusted. Tutto sommato, visto che due fattori non sono troppi, forse è meglio tenerli.

Riassumiamo i risultati di due degli indicatori principali nella seguente tabella:

Reg1: Multiple R-squared: 0.8464, Adjusted R-squared: 0.8055

Reg2: Multiple R-squared: 0.8464, Adjusted R-squared: 0.8176

Reg3: Multiple R-squared: 0.8464, Adjusted R-squared: 0.8283

Reg4: Multiple R-squared: 0.8198, Adjusted R-squared: 0.8098

Si vede che fino a Reg3 c’è sicuramente un miglioramento, mentre Reg4 è peggiore da questi punti di vista.

Multiple R-squared ha un “gomito” passando da 2 ad 1 fattore

Adjusted R-squared ha un massimo per 2 fattori.

Entrambi quindi indicano che 2 fattori è il punto dove fermarsi.

1. **PREVISIONE CON MODELLO DI REGRESSIONE**

Illustriamo ora come si usa un modello di regressione per fare previsioni. Supponiamo di avere solo la tabella di alcune regioni italiane. Ad esempio, togliamo FRVG e Basi dalla precedente (fingiamo di non averle):

B <- read.table ('clipboard', header=TRUE)

 PLIC SC SA.SC TD TMI

Piem 0.088 0.471 -0.707 -0.607 -0.395

Vaos -1.545 0.348 -0.642 -0.813 1.578

Lomb 0.202 1.397 -0.836 -0.790 -0.538

TrAA 0.677 0.435 -1.269 -0.966 -0.075

Vene 0.088 1.334 -1.210 -0.848 -0.497

Ligu 1.190 -0.247 0.470 -0.429 -0.354

EmRo 0.658 1.177 -1.315 -0.863 -0.347

Tosc 0.126 1.092 -0.795 -0.644 -1.355

Umbr -1.431 0.675 -0.140 -0.524 -1.287

Marc 0.278 1.090 -0.265 -0.702 -0.0006

Lazi 2.329 0.546 -0.080 -0.113 -0.014

Abru 0.335 -0.373 0.402 -0.456 0.040

Moli 0.658 -1.289 0.065 0.451 -1.151

Camp -1.811 -1.314 2.031 1.664 0.414

Pugl -0.766 -0.926 1.038 0.648 1.109

Cala -0.500 -1.727 1.571 2.153 0.632

Sici -0.918 -1.130 1.332 1.517 1.783

Sard 0.449 -0.403 0.717 1.285 -0.238

Osserviamo che si ottiene lo stesso risultato con

B0 = A[-6,]

B1=B0[-16,]

Ora vale:

> cor(B)

 PLIC SC SA.SC TD TMI

PLIC 1.0000000 0.2951897 -0.3749463 -0.3261446 -0.3936708

SC 0.2951897 1.0000000 -0.8696647 -0.8617002 -0.4473247

SA.SC -0.3749463 -0.8696647 1.0000000 0.9003586 0.4666751

TD -0.3261446 -0.8617002 0.9003586 1.0000000 0.4255334

TMI -0.3936708 -0.4473247 0.4666751 0.4255334 1.0000000

Si noti che certe correlazioni sono calate.

Questa volta non assegnamo i vettori PLIC = B[,1]

 Ecc, ma usiamo una nuova sintassi: B$PLIC. Ad esempio:

B$PLIC

Costruiamo ora i due modelli per TD già interessanti:

RegB3 = lm(B$TD~ B$SC+ B$SA.SC)

RegB4 = lm(B$TD~ B$SA.SC)

Vediamo i due summary: summary(RegB3)

> summary(RegB3)

dove vediamo che i risultati sono un po’ inferiori al modello con tutte le regioni. Ora passiamo alla previsione. Supponiamo di avere i valori di SC.SA ed SC di FRVG e Basi:

C <- read.table ('clipboard', header=TRUE)

 SC SA.SC

FrVG -0.005 -1.028

Basi -1.154 0.661

“Prevediamo” il valore di TD usando i due modelli e questi valori:

TD.RegB3.FrVG = RegB3$coef %\*%c(1, -0.005 , -1.028)

TD.RegB3.Basi = RegB3$coef %\*%c(1, -1.154 , 0.661)

c(TD.RegB3.FrVG, TD.RegB3.Basi)

-0.6317142 0.7902160

>

Confrontati coi valori veri:

 TD

FrVG -0.804

Basi 0.844

Il risultato non è così cattivo. Coglie il “segno” rispetto alla media nazionale ed anche l’ordine di grandezza della variazione.

Considerazioni conclusive. Il metodo ora esposto di eliminare uno o pochi individui, trovare il modello relativo a quelli mantenuti e testarlo sugli individui eliminati viene detto “cross-validation” e può essere usato per testare la bontà di un modello rispetto ad un altro (confrontare modelli). Ovviamente R^e ecc. sono già indicatori che servono a confrontare modelli; ad essi si aggiunge la cross-validation.