INTRODUZIONE. TABELLE, CORRELAZIONE E PLOT

Usare un foglio txt (o word o simile) di appoggio

Word ha il vantaggio che possiamo salvare un po' di risultati, anche grafici

Copiare su R il comando A <- read.table("clipboard") senza dare d’invio.

Aprire il file in rete “[tavola dati indicatori di benessere](http://users.dma.unipi.it/~flandoli/indicatori_benessere1.txt) “ e fare “copia”

Tornare su R e dare “invio”. La tavola è registrata nella tabella A. Provare a vederla scrivendo A e poi “invio”.

Ora, per nostra comodità espositiva, la incolliamo anche qui:

 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

FrVG 0.639 -0.005 -1.028 -0.804 -1.301

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

Basi -0.747 -1.154 0.661 0.844 2.001

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

Osservare la tabella con cor(A) e plot(A):

> 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

>



PRIMO ESERCIZIO DI REGRESSIONE

Costruire dei vettori con le colonne di A:

PLIC = A[,1]

SC = A[,2]

SC.SA = A[,3]

TD = A[,4]

TMI = A[,5]

Vogliamo un modello per TD basato sulle altre variabili. Eseguire la regressione

Reg1 = lm(TD~PLIC+SC+SC.SA+TMI)

Provare a chiedere Reg1. Chiedere poi summary(Reg1)

> summary(Reg1)

Call:

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

Residuals:

 Min 1Q Median 3Q Max

-0.82935 -0.20683 0.01375 0.23317 0.70087

Coefficients:

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

(Intercept) 0.0001030 0.0985596 0.001 0.99918

PLIC 0.0008371 0.1166154 0.007 0.99437

SC -0.3003680 0.1892197 -1.587 0.13327

SC.SA 0.6481730 0.1968985 3.292 0.00494 \*\*

TMI 0.0089889 0.1245278 0.072 0.94341

---

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

Residual standard error: 0.4408 on 15 degrees of freedom

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

F-statistic: 20.67 on 4 and 15 DF, p-value: 5.801e-06

>

Osservazioni generiche: buon R^2, buon p-value globale, almeno un fattore importante e forse due; non è un cattivo modello.

Iniziamo un procedimento di eliminazione. Osservando i p-value dei singoli fattori e la correlazione tra TD e le altre variabili, è chiaro che la peggiore è PLIC, seguita da TMI. Eseguiamo la regressione senza PLIC:

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

Poi chiediamo summary(Reg2)

> 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

>

R^2 non è diminuito. R^2 adjusted è migliorato. Persino i singoli p-values sono migliorati. Eseguiamo la regressione anche senza TMI:

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

Poi chiediamo summary(Reg3)

> 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.

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~SC.SA)

Poi chiediamo 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

>

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. Magari è migliore come sintesi (meno variabili ha dei vantaggi) e come p-value. La decisione se fermarsi a Reg3 o no è quindi soggettiva, ma guidata dai risultati.

SECONDO ESERCIZIO 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):

 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

Copiare su R il comando B <- read.table("clipboard") senza dare d’invio, poi copiare la tabella, tornare su R e dare invio. Controllare con B e invio. 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. Assegnamo i soliti vettori:

PLIC = B[,1]

SC = B[,2]

SC.SA = B[,3]

TD = B[,4]

TMI = B[,5]

Costruiamo ora i due modelli per TD già interessanti:

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

RegB4 = lm(TD~ SC.SA)

Vediamo i due summary: summary(RegB3)

> summary(RegB3)

Call:

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

Residuals:

 Min 1Q Median 3Q Max

-0.83264 -0.21625 -0.01661 0.28728 0.70275

Coefficients:

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

(Intercept) 0.006027 0.104901 0.057 0.9549

SC -0.323297 0.212030 -1.525 0.1481

SC.SA 0.621943 0.212608 2.925 0.0104 \*

---

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

Residual standard error: 0.4388 on 15 degrees of freedom

Multiple R-squared: 0.8361, Adjusted R-squared: 0.8142

F-statistic: 38.25 on 2 and 15 DF, p-value: 1.289e-06

> summary(RegB4)

Call:

lm(formula = TD ~ SC.SA)

Residuals:

Min 1Q Median 3Q Max

-0.83333 -0.25536 0.01932 0.31669 0.75351

Coefficients:

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

(Intercept) -0.02048 0.10765 -0.190 0.851

SC.SA 0.90387 0.10921 8.276 3.56e-07 \*\*\*

---

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

Residual standard error: 0.4566 on 16 degrees of freedom

Multiple R-squared: 0.8106, Adjusted R-squared: 0.7988

F-statistic: 68.5 on 1 and 16 DF, p-value: 3.56e-07

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:

 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$coefficient[2]\*(-0.005)+ RegB3$coefficient[3]\*(-1.028)+ RegB3$coefficient[1]

TD.RegB3.Basi = RegB3$coefficient[2]\*(-1.154)+ RegB3$coefficient[3]\*(0.661)+ RegB3$coefficient[1]

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

 SC SC

-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.

Vediamo il secondo modello, con solo un fattore:

TD.RegB4.FrVG = RegB4$coefficient[2]\*( -1.028)+ RegB4$coefficient[1]

TD.RegB4.Basi = RegB4$coefficient[2]\*( 0.661)+ RegB4$coefficient[1]

> c(TD.RegB4.FrVG, TD.RegB4.Basi)

 SC.SA SC.SA

-0.9496625 0.5769734

>

Non è male neppure questo.

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.