1. **ELIMINAZIONE DI FATTORI NELLA TABELLA “INDICATORI DI BENESSERE”**

(Conviene aprire due o più fogli R ed eseguire alcune simulazioni in parallelo, per confrontare meglio i risultati)

IB <- 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

Vogliamo un modello per TD basato sulle altre variabili. Il modello può avere uno scopo esplorativo, cioè per capire quali variabili (tra le altre 4) spieghi maggiormente la disoccupazione, oppure previsivo (nel momento in cui si esamini un caso nuovo in cui siano disponibili i valori delle altre 4 variabili e si voglia conoscere il TD del caso nuovo).

Eseguiamo innanzi tutto la regressione completa:

fit.1 = lm(TD~PLIC+SC+SA.SC+TMI, data=IB)

> summary(fit.1)

 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

SA.SC 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

Intanto, ecco alcune osservazioni generiche: buon R^2, buon p-value globale (questo però in genere vale sempre), almeno un fattore importante e forse due; non è un cattivo modello. L’importanza dei fattori si vede dai relativi p-values.

Iniziamo un procedimento di eliminazione dei fattori, cioè la ricerca di un modello più, semplice, con meno fattori, ma prestazioni paragonabili.

Per l’eliminazione di guarda primariamente ai p-value, tenendo d’occhio R^2, R^2 adjusted, e magari la matrice di correlazione

> round(cor(IB),2)

 PLIC SC SA.SC TD TMI

PLIC 1.00 0.32 -0.41 -0.37 -0.44

SC 0.32 1.00 -0.84 -0.85 -0.48

SA.SC -0.41 -0.84 1.00 0.91 0.51

TD -0.37 -0.85 0.91 1.00 0.49

TMI -0.44 -0.48 0.51 0.49 1.00

L’eliminazione non va mai eseguita a coppie o gruppi, come chiarito anche dalla teoria dei fattori allineati.

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:

fit.2 = lm(TD~SC+SA.SC+TMI, data=IB)

> summary(fit.2)

 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

SA.SC 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. Il modello è sicuramente migliore.

TMI può essere ulteriormente eliminato, non era un fattore allineato (un poco, ma è irrilevante).

fit.3 = lm(TD~SC+SA.SC, data = IB)

> summary(fit.3)

 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

SA.SC 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:

fit.4 = lm(TD~SC+SA.SC+0, data=IB)

> summary(fit.4)

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

SC -0.3020 0.1711 -1.765 0.0945 .

SA.SC 0.6511 0.1711 3.805 0.0013 \*\*

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

Residual standard error: 0.4024 on 18 degrees of freedom

Multiple R-squared: 0.8464, Adjusted R-squared: 0.8293

F-statistic: 49.58 on 2 and 18 DF, p-value: 4.766e-08

E’ migliorato ancora.

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:

fit.5 = lm(TD~SA.SC+0, data=IB)

>summary(fit.5)

 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.

1. **CONSIDERAZIONI AGGIUNTIVE SUGLI ALLINEAMENTI**

Abbiamo incontrato due esempi degli effetti degli allineamenti.

Nella regressione fit.1, PLIC ha coefficiente positivo. Ma nella matrice di correlazione, è correlato negativamente a TD. Questi due fatti sembrano intuitivamente in contraddizione (rigorosamente non lo sono, perché non cè nessun teorema che lega le due grandezze nella regressione lineare multipla; però in quella semplice avrebbero dovuto avere lo stesso segno).

La ragione è l’allineamento con TMI. Sappiamo che gli allineamenti possono provocare valori sballati dei coefficienti, incluso il loro segno. Questo è un esempio. Per verificare che è così, basta togliere TMI e vedere che il segno di PLIC si aggiusta:

summary(lm(TD~PLIC+SC+SA.SC, data=IB))

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

(Intercept) 0.0001046 0.0954465 0.001 0.99914

PLIC -0.0017242 0.1075771 -0.016 0.98741

SC -0.3021496 0.1816775 -1.663 0.11575

SA.SC 0.6502384 0.1886554 3.447 0.00332 \*\*

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: 29.38 on 3 and 16 DF, p-value: 9.623e-07

Un altro sintomo di allineamento si è visto genericamente dal miglioramento, anche se pur debole, dei vari p-value quando si eliminano fattori. Siccome i fattori eliminati erano debolmente allineati a quelli rimasti, il miglioramento è modesto; ma c’è.

Infine, come si vede dalla matrice di correlazione, due fattori sono molto allineati: SC ed SA.SC. Nonostante questo, il loro legame con TD è così forte da mostrare entrambi dei p-value più che decenti. Ma non appena ne togliamo uno dei due, il p-value dell’altro diventa molto migliore. Lo abbiamo visto in fit.5. Vediamolo anche togliendo SA.SC:

summary(lm(TD~ SC+0, data=IB))

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

SC -0.8500 0.1208 -7.038 1.06e-06 \*\*\*

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

Residual standard error: 0.5262 on 19 degrees of freedom

Multiple R-squared: 0.7228, Adjusted R-squared: 0.7082

F-statistic: 49.54 on 1 and 19 DF, p-value: 1.063e-06