Regressione logistica

Metodi Quantitativi per Economia, Finanza e Management

Esercitazione n°11

Consegna Lavoro di gruppo

- Scadenza per la consegna del lavoro di gruppo è fissata inderogabilmente per il giorno:

Venerdì 12 Gennaio 2018

- La consegna va effettuata entro le **ore 12** alla Sig.ra **Enrica Luezza** (Segreteria 4° Piano)

- Il materiale da consegnare consiste in:
  - stampa cartacea della presentazione in Power Point;
  - Chiavetta USB contenente:
    - questionario;
    - base dati in formato Excel;
    - Script di R;
    - presentazione Power Point.

N.B. Il supporto elettronico (chiavetta USB) non sarà restituito

Regressione logistica - Modello

Modello di regressione logistica

- si vuole modellare la relazione tra una variabile dipendente dicotomica (0-1) e un insieme di regressori che si ritiene influenzino la variabile dipendente

- la variabile dicotomica rappresenta presenza/assenza di un fenomeno (es. abbandono cliente, acquisto prodotto…)

- l'obiettivo è stimare l'equazione

$$\text{logit}\,(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

dove π:= Pr(Y=1 I X) è la probabilità che il fenomeno si verifichi

Regressione logistica – Esempio

DATA SET: banca_churn

	NAME	LABEL
1	cliente	Cliente
2	target	Target: abbandono
3	mavere	Numero movimenti avere
4	mdare	Numero movimenti dare
5	utenze	Numero utenze in c/c
6	pprod	Percentuale famiglie prodotti posseduti
7	flag_acc_sti	Accredito stipendio Y/N
8	mesi_bmov	Numero mesi bassa movimentazione ultimo semestre
9	PremiVita	Totale premi ass.ni Vita
10	NumAssDanni	Num ass.ni Danni
11	PremiDanni	Totale premi ass.ni Danni
12	AnzCliente	Anzianità cliente
13	NumAssVita	Num ass.ni Vita
14	eta	Età Cliente

Variabile Dipendente/Variabile Target:

0: non ha abbandonato la banca

1: ha abbandonato la banca

Obiettivo:

prevedere la probabilità di abbandono a partire da un insieme di variabili indipendenti e capire come queste ultime influenzano l'esito della variabile target

Regressione logistica – Esempio

Qual è la percentuale di clienti che ha abbandonato la banca?

```
banca_churn=read.csv('banca_churn.csv', header=TRUE)
library(descr)
freq(banca_churn$target)
```

```
> freq(banca_churn$target)
banca_churn$target
          Frequency Percent
0             31281   81.97
1              6882   18.03
Total         38163  100.00
```

glm – Sintassi

Modello di regressione logistica – k regressori

In un modello di regressione logistica la variabile dipendente (Y) **DEVE** essere binomiale, ossia una variabile che assume il valore 0 o 1.

```
mylogit <- glm(variabile_dipendente_0_1 ~
regressore1 + regressore2 + … + regressoreK, data =
dataset_input, family = "binomial")
```

Indica la distribuzione della variabile risposta

N.B: le variabili continue, Es. 1.520,40 devono essere scritte nel file .csv come 1,520.40 (formato numerico americano).

glm – Esempio

```
mylogit <- glm(target ~ mesi_bmov + pprod + utenze + mdare
+ mavere + flag_acc_sti + eta + NumAssVita + NumAssDanni +
AnzCliente, data = banca_churn, family = "binomial")
```

```
summary(mylogit)
```

```
Call:
glm(formula = target ~ mesi_bmov + pprod + utenze + mdare + mavere +
    flag_acc_sti + eta + PremiVita + PremiDanni + NumAssVita +
    NumAssDanni + AnzCliente, family = "binomial", data = banca_churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1562  -0.5015  -0.2818  -0.1156   7.1341

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.188e-01  7.264e-02   4.389 1.14e-05 ***
mesi_bmov     4.457e-01  8.476e-03  52.587  < 2e-16 ***
pprod        -5.326e+00  1.987e-01 -26.807  < 2e-16 ***
utenze       -5.855e-02  1.337e-02  -4.380 1.19e-05 ***
mdare        -4.413e-02  3.174e-03 -13.906  < 2e-16 ***
mavere       -2.836e-01  1.446e-02 -19.609  < 2e-16 ***
flag_acc_sti -8.584e-01  4.965e-02 -17.287  < 2e-16 ***
eta           7.280e-04  1.101e-03   0.661    0.508
PremiVita     2.229e-05  3.007e-05   0.741    0.458
PremiDanni   -6.626e-05  5.646e-05  -1.174    0.241
NumAssVita    3.935e-02  3.808e-02   1.033    0.301
NumAssDanni   9.250e-03  2.353e-02   0.393    0.694
AnzCliente    1.647e-03  3.676e-03   0.448    0.654
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 36018  on 38162  degrees of freedom
Residual deviance: 24396  on 38150  degrees of freedom
AIC: 24422
```

Le stime dei coefficienti non sono tutte significative, svolgiamo la stepwise

glm – Esempio, stepwise

```
a=step(mylogit, direction='both')

summary(a)
```

```
> summary(a)

Call:
glm(formula = target ~ mesi_bmov + pprod + utenze + mdare + mavere +
    flag_acc_sti, family = "binomial", data = banca_churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1496  -0.5024  -0.2823  -0.1159   7.1298

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.377089   0.044533   8.468  < 2e-16 ***
mesi_bmov    0.445552   0.008473  52.583  < 2e-16 ***
pprod       -5.323200   0.198603 -26.803  < 2e-16 ***
utenze      -0.058421   0.013363  -4.372 1.23e-05 ***
mdare       -0.044127   0.003172 -13.910  < 2e-16 ***
mavere      -0.283493   0.014462 -19.603  < 2e-16 ***
flag_acc_sti -0.858018   0.049643 -17.284  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 36018  on 38162  degrees of freedom
Residual deviance: 24399  on 38156  degrees of freedom
AIC: 24413

Number of Fisher Scoring iterations: 6
```

glm – Interpretazione dei coefficienti

```
> summary(a)

Call:
glm(formula = target ~ mesi_bmov + pprod + utenze + mdare + mavere +
    flag_acc_sti, family = "binomial", data = banca_churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1496  -0.5024  -0.2823  -0.1159   7.1298

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.377089   0.044533   8.468  < 2e-16 ***
mesi_bmov      0.445552   0.008473  52.583  < 2e-16 ***
pprod         -5.323200   0.198603 -26.803  < 2e-16 ***
utenze        -0.058421   0.013363  -4.372 1.23e-05 ***
mdare         -0.044127   0.003172 -13.910  < 2e-16 ***
mavere        -0.283493   0.014462 -19.603  < 2e-16 ***
flag_acc_sti -0.858018   0.049643 -17.284  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 36018  on 38162  degrees of freedom
Residual deviance: 24399  on 38156  degrees of freedom
AIC: 24413

Number of Fisher Scoring iterations: 6
```

Stime dei parametri

Stime standardizzate

Osservo il segno e l'importanza

```
> lm.beta(a)
   mesi_bmov        pprod       utenze        mdare       mavere flag_acc_sti
   1.8191874   -1.6999811   -0.2811542   -1.1388725   -2.2038170   -1.0205245
```

Stime odds-ratio

```
exp(a$coefficient)
```

```
> exp(a$coefficient)
 (Intercept)    mesi_bmov        pprod       utenze        mdare       mavere
 1.458033898  1.561351974  0.004877121  0.943253097  0.956832705  0.753148021
flag_acc_sti
 0.424001809
```

Stime odds-ratio, interpretazione

Valutazione bontà del modello

Valutazione bontà del modello

1. **Percentuale di Concordant** → valuta la capacità del modello di stimare la probabilità che il fenomeno si verifichi (quanto più la percentuale è alta tanto migliore è il modello)

2. **Test di significatività congiunta dei coefficienti (Likelihood ratio test/score test/Wald test)** → OK p-value inferiori al livello di significatività fissato

   →equivalenti al test F nella regressione lineare (valuta la capacità esplicativa del modello)

3. **Test di significatività dei singoli coefficienti (Wald Chi_square test)** → OK p-value inferiori al livello di significatività fissato

   → equivalente al test t nella regressione lineare (valuta la significatività dei singoli coefficienti = la rilevanza dei corrispondenti regressori nella spiegazione della variabile dipendente)

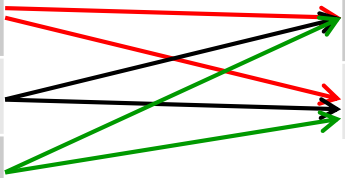
Percentuale di concordant: come si calcola?(1/2)

Dataset con Y=1

VALORE VARIABILE DIPENDENTE ORIGINALE	SCORE
1	0.3
1	0.9
1	0.7

Dataset con Y=0

VALORE VARIABILE DIPENDENTE ORIGINALE	SCORE
0	0.5
0	0.8



1. Si divide la tabella iniziale in due tabelle: nella prima si trovano tutte le osservazioni la cui variabile dipendente assume valore 1, nell'altra quelle la cui variabile dipendente assume valore 0.
2. Si confronta ogni osservazione della prima tabella con ognuna delle osservazione nella seconda tabella (si formeranno quindi n*m **coppie**, dove n=osservazioni tabella 1, m=osservazioni tabella 2)
3. Si assegnano I seguenti punteggi:
   CONCORDANTI=1 se **score della prima tabella > score seconda tabella ,** altrimenti 0
   DISCORDANTI=1 se **score della prima tabella < score seconda tabella ,** altrimenti
   TIED= 1 se **score della prima tabella = score seconda tabella,** altrimenti 0
4. La **percentuale di concordant** è calcolata sommando i **CONCORDANTI** e dividendoli per il numero totale delle coppie (in modo analogo la percentuale di **discordant e tied**)

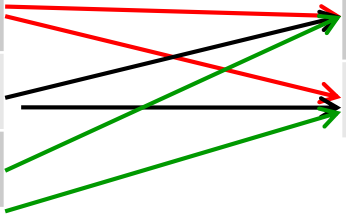
Percentuale di concordant: come si calcola?(2/2)

Dataset con Y=1

VALORE VARIABILE DIPENDENTE ORIGINALE	SCORE
1	0.3
1	0.9
1	0.7

Dataset con Y=0

VALORE VARIABILE DIPENDENTE ORIGINALE	SCORE
0	0.5
0	0.8



Nell'esempio riportato quindi:

Numero di coppie: 3x2=6

Punteggi Concordanti (per ogni coppia):

0 0 1 1 1 0

Punteggi Discordanti (per ogni coppia):

1 1 0 0 0 1

Tied:

00 0 0 0 0 0

Percentuale di concordanti:

(0 +0 + 1+ 1+ 1 + 0)/6 = 3/6= 0.5

Percentuale disconcordanti:

(1 +1 + 0+ 0+ 0 + 1)/6 = 3/6 = 0.5

Tied=

0

Percentuale di Concordant - Sintassi

Per calcolare la percentuale di Concordant bisogna richiamare la funzione

```
> ########################################funzione per calcolare concordanti e discordanti#####################################
>
> # Function to calculate concordance and discordance
> CalculateConcordance <- function (myMod){
+ fitted <- data.frame (cbind (myMod$y, myMod$fitted.values)) # actuals and fitted
+ colnames(fitted) <- c('response','score') # rename columns
+ ones <- fitted[fitted$response==1, ] # Subset ones
+
+ zeros <- fitted[fitted$response==0, ] # Subsetzeros
+
+ totalPairs <- nrow (ones) * nrow (zeros) # calculate total number of pairs to check
+
+ conc <- sum (c (vapply (ones$score, function(x) {((x > zeros$score))}, FUN.VALUE=logical(nrow(zeros)))))
+
+ disc <- totalPairs - conc
+
+ # Calc concordance, discordance and ties
+
+ concordance <- round(conc/totalPairs,digit=4)
+
+ discordance <- round(disc/totalPairs,digit=4)
+
+ tiesPercent <- round((1-concordance-discordance),digit=4)
+
+ return(list("Concordance"=concordance, "Discordance"=discordance,"Tied"=tiesPercent, "Pairs"=totalPairs))
+ }
> ########################################fine funzione per calcolare concordanti e discordanti#################################
>
```

Eseguire la funzione.

`CalculateConcordance(`*nome del modello*`)`

Percentuale di Concordant - Output

`CalculateConcordance(`*`mylogit`*`)`

```
> CalculateConcordance(mylogit)
$Concordance
[1] 0.8691

$Discordance
[1] 0.1309

$Tied
[1] 0

$Pairs
[1] 215275842
```

Percentuale di Concordant → valuta la capacità del modello di stimare la probabilità che il fenomeno si verifichi (quanto più la percentuale è alta tanto migliore è il modello)

Test di significatività congiunta dei coefficienti

Test di significatività congiunta dei coefficienti: Wald test

$$H_0 : \beta_1 = ... = \beta_p = 0$$
$$H_1 : \text{almeno un } \beta_j \neq 0$$

Il Wald Test è equivalente al test F nella regressione lineare: valuta la capacità esplicativa del modello

Per calcolare il test di Wald in R bisogna scaricare un pacchetto:

```
library(lmtest)
```

```
waldtest(nome_modello_glm)
```

Test di significatività congiunta dei coefficienti

Test di significatività congiunta dei coefficienti: Wald test

$$H_0 : \beta_1 = \ldots = \beta_p = 0$$

$$H_1 : \text{almeno un } \beta_j \neq 0$$

Il Wald Test è equivalente al test F nella regressione lineare: valuta la capacità esplicativa del modello

```
waldtest(a)
```

```
> waldtest(a)
Wald test

Model 1: target ~ mesi_bmov + pprod + utenze + mdare + mavere + flag_acc_sti
Model 2: target ~ 1
  Res.Df Df      F    Pr(>F)
1  38156
2  38162 -6 1153.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test di significatività dei singoli coefficienti

Test di significatività per i singoli coefficienti

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.377089   0.044533   8.468  < 2e-16 ***
mesi_bmov    0.445552   0.008473  52.583  < 2e-16 ***
pprod       -5.323200   0.198603 -26.803  < 2e-16 ***
utenze      -0.058421   0.013363  -4.372 1.23e-05 ***
mdare       -0.044127   0.003172 -13.910  < 2e-16 ***
mavere      -0.283493   0.014462 -19.603  < 2e-16 ***
flag_acc_sti -0.858018   0.049643 -17.284  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Wald Chi_square test**

valuta la significatività dei singoli coefficienti, ossia la rilevanza dei corrispondenti regressori nella spiegazione della variabile dipendente (equivalente al test t nella regressione lineare)

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Multicollinearità

Per valutare la presenza di multicollinearità tra i regressori, si usa l'indicatore VIF → specificare solo i regressori significativi

Per calcolare l' indicatore VIF, è necessario scaricare un pacchetto:

usdm e richiamarlo.

Successivamente si potrà usare la funzione vif.

```
library(usdm)
```

```
vif(nome_subset_input)
```

Multicollinearità

```
banca_parametri<-banca_churn[ ,c("mesi_bmov", "pprod",
"utenze", "mdare", "mavere", "flag_acc_sti" )]

library(usdm)

v=vif(banca_parametri)
```

```
> v<-vif(banca_parametri)
> v
     Variables      VIF
1    mesi_bmov 1.163165
2        pprod 1.507258
3       utenze 1.513167
4        mdare 1.540348
5       mavere 1.281604
6 flag_acc_sti 1.178926
```

RICORDATE:

Un VIF = 1 significa che quella variabile non e' coinvolta in nessuna situazione di multicollinearità. VIF superiore a 1,3 indica che la presenza di almeno un po' di multicollinearità

Multicollinearità

Per risolvere il problema della multicollinearità, è necessario ricorrere ad una delle seguenti azioni:

- rimuovere le variabili indipendenti affette da multicollinearità;
- mantenere nel modello una sola variabile tra quelle indipendenti affette da multicollinearità;
- eseguire una analisi fattoriale su TUTTE le variabili indipendenti di partenza (l'esito della stepwise potrebbe essere stato influenzato dalla presenza di multicollinearità);

Esempio di risoluzione multicollinearità

Multicollinearità

Esempio di risoluzione multicollinearità

```
banca_subset<-banca_churn[ ,c("mesi_bmov", "pprod",
"utenze", "mdare", "mavere","flag_acc_sti", "eta",
"PremiVita", "PremiDanni", "NumAssVita", "NumAssDanni",
"AnzCliente" )]

fit=princomp(banca_subset, cor=TRUE)

summary(fit)

library("factoextra")

Eig.val <- get_eigenvalue(fit)

Eig.val

plot(fit, type='lines')
```

Multicollinearità – risoluzione (1/6)

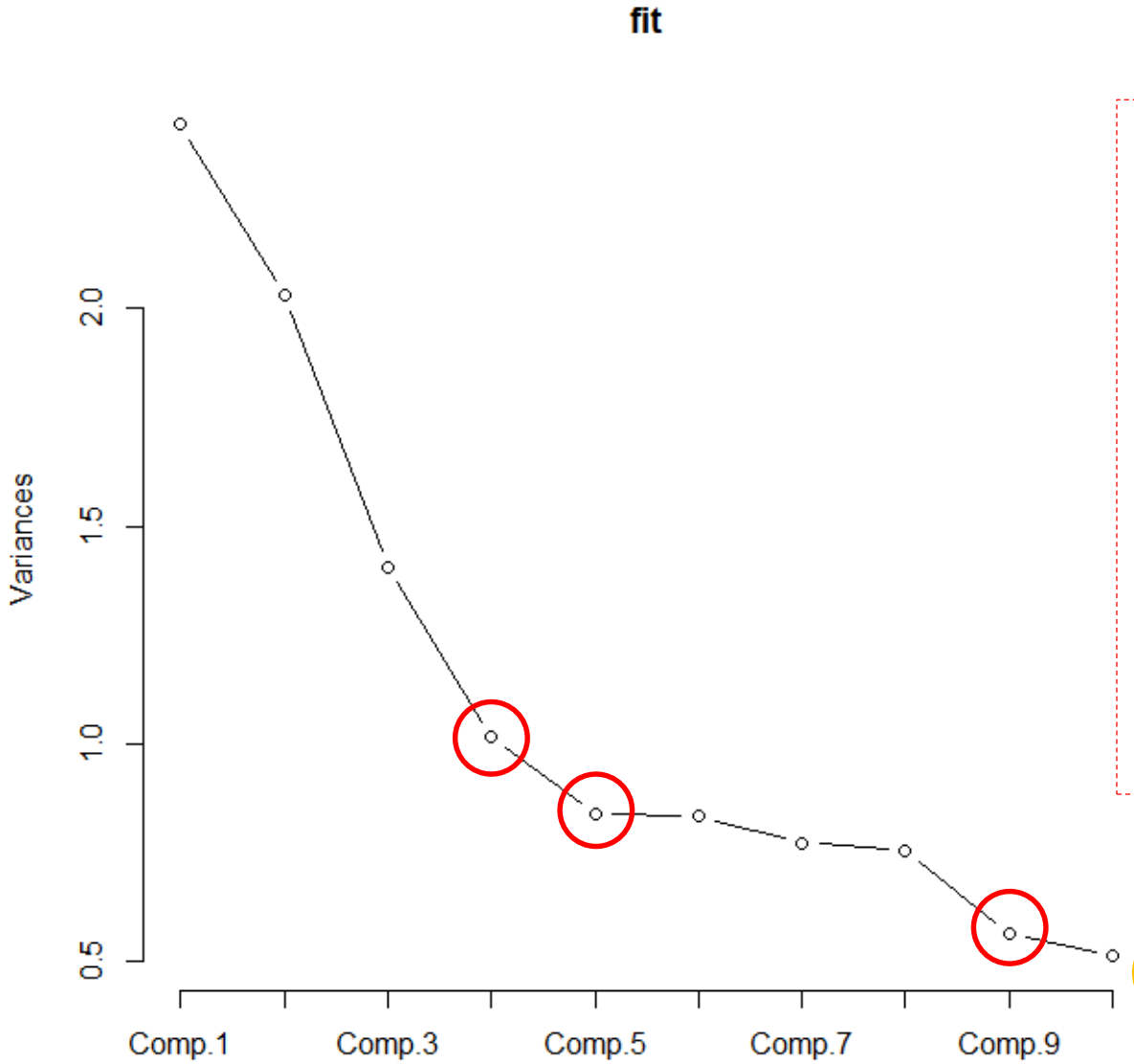
```
> Eig.val
       eigenvalue variance.percent cumulative.variance.percent
Dim.1   2.4222085        20.185071                    20.18507
Dim.2   2.0310069        16.925057                    37.11013
Dim.3   1.4039016        11.699180                    48.80931
Dim.4   1.0148432         8.457027                    57.26634
Dim.5   0.8397478         6.997899                    64.26423
Dim.6   0.8331757         6.943130                    71.20736
Dim.7   0.7732057         6.443381                    77.65075
Dim.8   0.7537958         6.281631                    83.93238
Dim.9   0.5660571         4.717142                    88.64952
Dim.10  0.5113684         4.261403                    92.91092
Dim.11  0.4450695         3.708913                    96.61983
Dim.12  0.4056198         3.380165                   100.00000
```

La regola degli autovalori > 1 suggerisce di prendere in considerazione **4 fattori**

Tale soluzione spiega il 57% della varianza originaria

%varianza spiegata >60%

Multicollinearità – risoluzione (2/6)



Lo scree plot mostra un gomito accennato in corrispondenza del 4 fattore, e due ulteriori gomiti in corrispondenza del fattore 5 e del fattore 9.

- 4 fattori → già selezionata
- 9 fattori → n° fattori elevato rispetto a n° variabili originarie
- **5 fattori → % varianza originaria spiegata = 64 %, adeguata**

N°fattori = circa 1/3 variabili originali → circa 4 fattori

Multicollinearità – risoluzione (3/6)

Confronto tra le comunalità delle soluzioni a 4 e a 5 fattori.

```
principal(banca_subset,nfactors=4,residuals=FALSE,rotate="none")$communality
principal(banca_subset,nfactors=5,residuals=FALSE,rotate="none")$communality
principal(banca_subset,nfactors=6,residuals=FALSE,rotate="none")$communality
```

	4 fattori	5 fattori	6 fattori
mesi_bmov	0,29	0,30	0,86
pprod	0,58	0,58	0,58
utenze	0,56	0,56	0,57
mdare	0,74	0,74	0,74
mavere	0,41	0,43	0,63
flag_acc_sti	0,86	0,86	0,90
eta	0,41	0,86	0,89
PremiVita	0,49	0,81	0,83
PremiDanni	0,63	0,64	0,64
NumAssVita	0,61	0,61	0,61
NumAssDanni	0,73	0,73	0,73
AnzCliente	0,57	0,57	0,57

La variabile «mesi_bmov» non risulta adeguatamente spiegata dalle due soluzioni → analizziamo anche la soluzione a 6 fattori (71% varianza spiegata)

La soluzione a 6 fattori spiega adeguatamente la variabile «mesi_bmov», ma porterebbe ad un numero di componenti principali troppo elevato rispetto alle variabili originarie.

Decidiamo di proseguire con un tentativo di **interpretazione per la soluzione a 5 fattori**.

Multicollinearità – risoluzione (4/6)

Rotazione dei fattori con il metodo Varimax ed interpretazione.

```
principal(banca_subset,nfactors=5,residuals=FALSE,rotate="varimax")
```

```
> y<-principal(banca_subset,nfactors=5,residuals=FALSE,rotate="varimax")
> print(y$loadings,sort=T)

Loadings:
           RC1    RC2    RC3    RC4    RC5
pprod       0.716                 0.252
utenze      0.713                 0.232
mdare       0.802                -0.314
mavere      0.633                       -0.129
PremiDanni         0.799
NumAssVita         0.762                 0.139
NumAssDanni        0.854
PremiVita                 0.898
AnzCliente        -0.102 -0.620         0.415
flag_acc_sti 0.187                0.910
eta                0.154                0.912
mesi_bmov  -0.500                -0.188 -0.108
```

Ipotesi di intepretazione:

Movimentazioni conto corrente

Prodotti assicurativi

???

Multicollinearità – risoluzione (5/6)

L'interpretazione della soluzione selezionata non è soddisfacente...

NOTA BENE!!!

In ambito di risoluzione della multicollinearità, l'analisi fattoriale non ha l'obiettivo primario di sintetizzare un elevato numero di variabili correlate tra loro.

⇩

L'esigenza di parsimonia nella selezione dei fattori diventa meno stringente!

COSA FARE?

Nell'ottica di sanare il problema della multicollinearità: se l'interpretazione della soluzione ottimale, suggerita dai criteri pratici e dal confronto delle comunalità, non è convincente, possiamo provare ad ammettere un numero più elevato di fattori, purchè vi sia un guadagno in termini di interpretabilità.

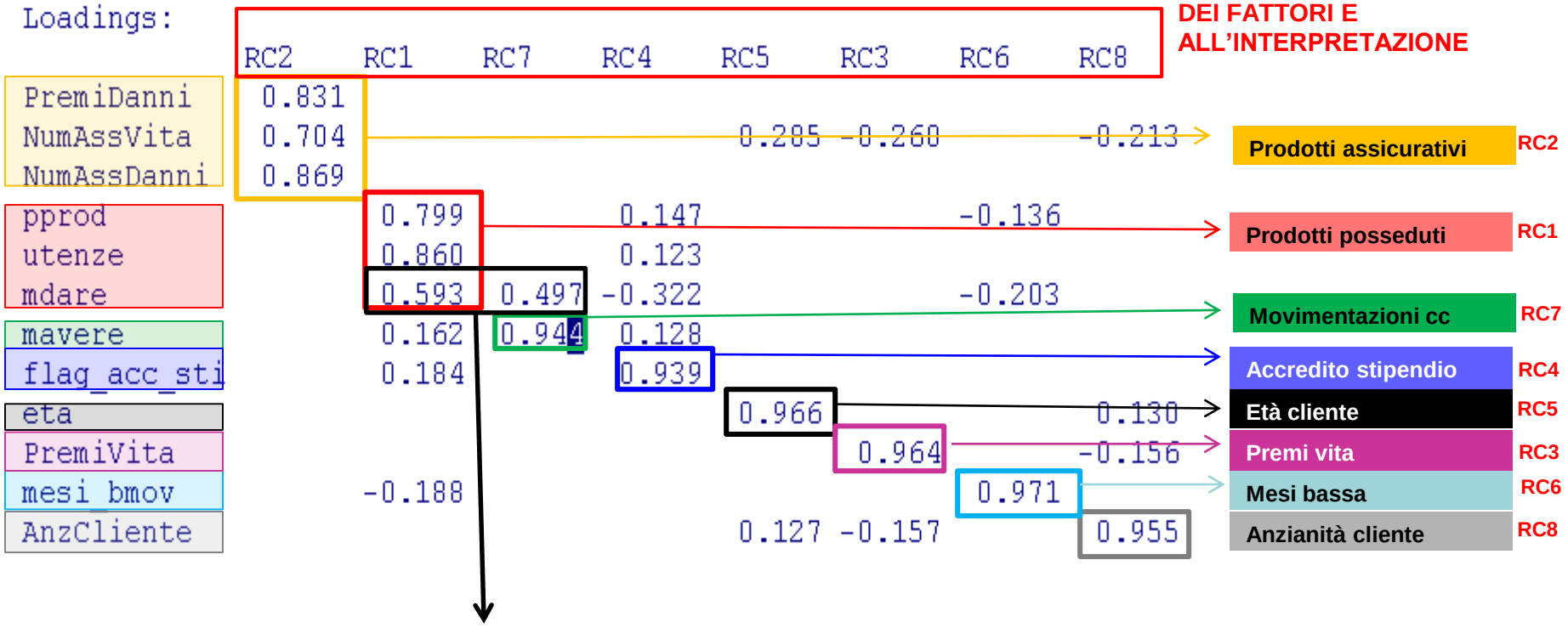
Multicollinearità – risoluzione (6/6)

Proviamo a rieseguire lo step di interpretazione aumentando di volta in volta il numero di fattori considerati (nell'esempio: 6 fattori, 7 fattori, ecc). Ci arrestiamo quando la soluzione analizzata fornisce una interpretazione soddisfacente.

```
y2<-principal(banca_subset,nfactors=8,residuals=FALSE,rotate="varimax")

print(y2$loadings,sort=T)
```

ATTENZIONE ALL'ORDINE DEI FATTORI E ALL'INTERPRETAZIONE



N.B.: la variabile «mdare» ha correlazioni simili con RC1 e RC7 → è opportuno tener conto del significato della variabile anche nell'interpretazione di RC7!

Stima modello

Stima del modello considerando i fattori estratti come variabili indipendenti.

```
mylogit_factor<-glm(target~RC1+RC2+RC3+RC4+RC5+RC6+RC7+RC8
                          ,data=banca_scored,family="binomial")
a2<-step(mylogit_factor,direction="both")
summary(a2)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37359    0.02504  -94.79   <2e-16 ***
RC1         -1.25835    0.02303  -54.64   <2e-16 ***
RC4         -0.47925    0.02268  -21.13   <2e-16 ***
RC6          0.91207    0.01308   69.72   <2e-16 ***
RC7         -1.15231    0.03548  -32.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

RC1 → prodotti posseduti
RC7 → movimentazioni conto corrente
RC4 → flag accredito stipendio
RC6 → numero mesi bassa movimentazione ultimo semestre

Bontà del modello

Valutazione della bontà del modello

1. WALD TEST

```
> waldtest(a2)
Wald test

Model 1: target ~ RC1 + RC4 + RC6 + RC7
Model 2: target ~ 1
  Res.Df Df      F    Pr(>F)
1  38158
2  38162 -4 1709.2 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. PERCENTUALE CONCORDANTI

```
> CalculateConcordance(mylogit_factor)
$Concordance
[1] 0.8667

$Discordance
[1] 0.1333

$Tied
[1] 0

$Pairs
[1] 215275842
```

Multicollinearità

Verifica della presenza di multicollinearità per il nuovo modello stimato (solo i regressori significativi).

```
banca_parametri_factor<-banca_scored[,c("RC1","RC4","RC6","RC7")]
vif(banca_parametri_factor)
```

```
> vif(banca_parametri_factor)
  Variables      VIF
1       RC1 1.001815
2       RC4 1.000625
3       RC6 1.000340
4       RC7 1.002245
```

RICORDATE:

Un VIF = 1 significa che quella variabile non e' coinvolta in nessuna situazione di multicollinearità. VIF superiore a 1,3 indica che la presenza di almeno un po' di multicollinearità

E' stato risolto il probelma della Multicollinearità

Interpretazione dei Coefficienti (1/3)

Importanza dei regressori – coefficienti standardizzati

```
> lm.beta(a2)
       RC1        RC4        RC6        RC7
-3.272952 -1.246537  2.372287 -2.997149
```

Si ordinano i regressori in modo decrescente rispetto al valore assoluto del coefficiente standardizzato.

Il fattore RC1 (prodotti posseduti) è il regressore maggiormente influente nel modello. Seguono in termini di importanza il fattore RC7, il fattore RC6 e il fattore RC4.

Interpretazione dei Coefficienti (2/3)

Analisi del segno dei coefficienti standardizzati

```
> lm.beta(a2)
       RC1        RC4        RC6        RC7
-3.272952 -1.246537  2.372287 -2.997149
```

- Più è elevato il numero di prodotti posseduti, più diminuisce la probabilità di abbandono (coeff. std. RC1= -3.272 segno negativo)
- Più è alta la movimentazione di C/C, più diminuisce la probabilità di abbandono (coeff. std. RC7= -2.997149 segno negativo)
- La presenza dell'accredito dello stipendio in C/C diminuisce la probabilità di abbandono (coeff. std. RC4= -1.2465 segno negativo)
- Più aumentano i mesi di bassa movimentazione nell'ultimo semestre, più aumenta la probabilità di abbandono (coeff. std. RC6= 2.372287 segno positivo)

Interpretazione dei Coefficienti (3/3)

Interpretazione dei regressori – stime odds-ratio

Exp(a2$coefficient)

```
> exp(a2$coefficient)
(Intercept)         RC1         RC4         RC6         RC7
  0.0931461   0.2841233   0.6192449   2.4894698   0.3159064
```

REGOLA: poniamo soglia 1 e verifichiamo se gli ODDS-RATIO sono sopra o sotto soglia. Calcolare la differenza tra la stima odds-ratio e soglia 1 per interpretare i regressori.

All'aumentare dei mesi bassa movimentazione ultimo semestre (RC6), aumenta la probabilità che il cliente abbandoni la banca del 148% (2 volte e mezzo -> 2.48-1=1.48).

I clienti che accreditano lo stipendio (RC4), hanno circa il 40% di probabilità in meno di abbandonare la banca rispetto a chi non accredita lo stipendio.

Regressione logistica – Passi da fare

1) Individuare la variabile oggetto di analisi (variabile dipendente dicotomica (0/1)) e i potenziali regressori (variabili quantitative o dummy).
2) Stimare un modello di regressione logistica utilizzando il metodo di selezione automatica STEPWISE per selezionare le variabili.
3) Valutare:
    I. la bontà del modello (*percentuale di Concordant*);
    II. la significatività congiunta dei coefficienti (*Wald test*);
    III. la significatività dei singoli coefficienti stimati (*Wald Chi-square test*).
4) Valutare la presenza di multicollinearità tra i regressori

Regressione logistica – Passi da fare

5) Nel caso di multicollinearità, provvedere alla risoluzione del problema tramite una delle seguenti opzioni:
    - rimuovere le variabili indipendenti affette da multicollinearità;
    - mantenere nel modello una sola variabile tra quelle indipendenti affette da multicollinearità;
    - analisi fattoriale su tutte le variabili indipendenti di partenza.
6) Rieseguire gli step 2-3-4-5 fino ad individuare il modello finale.
7) Interpretare i coefficienti standardizzati:
    I. stabilire tra i regressori un ordine di importanza nella spiegazione della variabile target;
    II. valutare la direzione dell'impatto di ogni regressore sulla variabile target, tramite analisi del segno dei coefficienti.
8) Interpretazione odds-ratio.