Dall'Analisi Fattoriale alla Regressione Lineare

Metodi Quantitativi per Economia, Finanza e Management

Esercitazione n° 10

Consegna Lavoro di gruppo

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

Giovedì 10 Gennaio 2019

- 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 R;
    - presentazione Power Point

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

Analisi Fattoriale

Tecnica di analisi multivariata

Quando si utilizza?

- Nel caso di un elevato numero di variabili quantitative, tra loro correlate (linearmente).
- NB: in contesti applicativi, è usata anche con variabili qualitative ordinali che esprimono scale di preferenza numeriche (scale di punteggi).

Perché si utilizza?

- Informazione condivisa tra le variabili correlate → è ridondante utilizzarle tutte
- Informazione dispersa tra le variabili → possibilità che le variabili, utilizzate singolarmente, siano poco esplicative

Analisi Fattoriale

OBIETTIVO

Sintetizzare le variabili originarie in un numero inferiore di variabili, dette fattori "LATENTI"

FATTORI LATENTI:

- concetti non direttamente misurabili

  Esempio: la qualità della vita non è direttamente misurabile. Sono misurabili invece: il tasso di disoccupazione, tasso di aree verdi, tasso di inquinamento, aspettativa di vita…

- caratterizzati da una maggior facilità interpretativa

- spiegano «buona parte» della variabilità originaria, ovvero del contributo informativo delle variabili di partenza

Processo di analisi

Identificazione *p* variabili di partenza (variabili quantitative o scale di punteggio)

Selezione di *k* fattori

(dove *k* <*p*)

- **Calcolo le p componenti principali.**
- **Utilizzo di alcuni criteri per la *selezione dei possibili valori di k* (è possibile identificare più valori di k adeguati)**
- ***Confronto* tra le possibili soluzioni identificate (confronto delle comunalità)**
- ***Verifica dell'interpretabilità* della soluzione scelta ed eventuale indagine di una soluzione differente**

Interpretazione della soluzione finale

Analisi fattoriale (1/2)

STEP 1: scegliere quanti fattori considerare (scelta di varie soluzioni)

- la regola autovalori > 1
- lettura dello SCREE PLOT
- Circa 1/3 delle variabili originarie
- Variabilità spiegata > 60%

```
Nome_1 = princomp(nome_subset, cor=TRUE)

get_eigenvalue(nome_1)

plot(nome_1, type='lines')
```

STEP 2: confrontare le soluzioni scelte

- cumunalità finali

```
Nome_2= principal(nome_subset, residuals=FALSE, nfactors = num_fattori,
rotate ='none')
Nome_2$communality
```

Analisi fattoriale (2/2)

STEP 3: una volta scelta la soluzione finale

- ruotare i fattori
- interpretare i fattori
- salvare il data set con i fattori e rinominarli

```
nome_3= principal(nome_subset, residuals=FALSE, nfactors = num_fattori,
rotate ='varimax', score=TRUE)

nome_3$loadings

fa.diagram(nome_3)

Nome_4=cbind(dataset_originale, nome_2$scores)
```

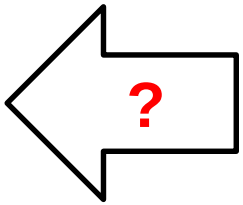
STEP 4: se l'interpretazione non è soddisfacente ripetere lo step n°3 variando metodo di rotazione o provando un'altra soluzione.

Modello di Regressione Lineare

L'analisi della regressione lineare è una metodologia asimmetrica che si basa sull'ipotesi dell'esistenza di una relazione di tipo **causa-effetto** tra una o più variabili indipendenti (o esplicative, X_i) e la variabile dipendente (Y).

Y

Variabile «target»: rappresenta un fenomeno di interesse (variabile quantitativa continua)



$X_1, X_2, \ldots, X_p$

Variabili che si ritiene possano influenzare Y

OBIETTIVO:

Individuare quali variabili tra X1,…,Xp (variabili «indipendenti») influenzano la variabile Y (variabile «dipendente») e come la influenzano

Regressione lineare (1/3)

1. Individuazione variabili dipendente e regressori
2. Trasformazione di eventuali variabili qualitative in dummy
3. Stimare un modello di regressione lineare utilizzando la procedura automatica di selezione delle variabili (stepwise)
4. Valutare la bontà del modello (R-square, Test F, Test t)
5. Se la procedura stepwise non ha prodotto tutte stime significative, provare a stimare un modello di regressione lineare con i soli parametri le cui stime sono significative.
   Tornare al punto 4, poi al punto 6.

Regressione lineare (2/3)

6. Verificare la presenza di multicollinearità (se i regressori del modello sono i fattori di un'analisi fattoriale non è necessario perchè risultano non correlati per costruzione → tutti i VIFj =1)

   - ✓ Se si è in presenza di multicollinearità: azioni per eliminarla (provare ad eliminare una variabile e verificare che il nuovo modello abbia VIF accettabili; oppure effettuare analisi fattoriale sui regressori) e ripetere i punti 3, 4
   - ✓ In assenza di multicollinearità: passare al punto 7

7. Verificare l'impatto dei regressori nella spiegazione del fenomeno (ordinarli usando il valore assoluto dei coefficienti standardizzati e controllare il segno dei coefficienti)

Regressione lineare (2/3)

8. Interpretazione del coefficienti standardizzati

Se il regressore3 *standardizzato* aumenta di una unità allora la variabile dipendente *standardizzata* diminuisce di 0,31

Se il regressore3 *standardizzato* diminuisce di una unità allora la variabile dipendente *standardizzata* aumenta di 0,31

N.B.:attenzione al segno del coefficiente!!

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	1.71	0.283	6.03	<.0001	0
regressore 1	1	0.12	0.032	3.77	<.0001	0.19
regressore 2	1	0.08	0.026	2.99	<.0001	0.13
regressore 3	1	-0.22	0.034	6.29	<.0001	-0.31
regressore 4	1	0.18	0.037	4.81	<.0001	0.26

Regressione lineare (3/3)
Sintassi

- Stimare un modello di regressione lineare

```
Nome_dataset_lm = lm (variabile_dipendente ~
   variabile_indipendente, data=dataset_input)
```

- Stimare un modello di regressione lineare con il metodo stepwise

```
nome_modello_lm = step(nome_dataset_lm, direction='both')
```

- Visualizzare il modello e valutarne la bontà (R-Quadro, test F, test t)

```
summary(nome_modello_lm)
```

- Calcolare le stime standardizzate dei coefficienti

```
lm.beta(nome_modello_lm)
```

- Verifica presenza multicollinearità

```
vif(nome_subset_input)
```

Esercizio

Il dataset ct_telefonia.csv contiene i dati di 126.761 clienti di una compagnia telefonica e 25 variabili quantitative.

#	Variable	Descrizione
1	AMMONT_RICARICA_BONUS	Ammontare ricariche bonus
2	AMMONT_RICARICA_PAG	Ammontare ricariche pagate dal cliente
3	AMMONT_RICARICA_PAG_LOTTO	Ammontare ricariche effettuate tramite circuito lotto sisal
4	AMMONT_RICARICA_RICORRENTE	Ammontare ricariche ricorrenti
5	ANZIANITA_SIM	Anzianità della SIM espressa in mesi
6	CONTATTI_INBOUND	Numero di volte in cui il cliente ha contattao il call center negli ultimi 6 mesi
7	CONTATTI_OUTBOUND	Numero di volte in cui il call center ha contattato il cliente negli ultmi 6 mesi (per campagna commerciale)
8	D_OPZ_ESTERO	Variabile che indica se è attiva, disattiva o dismessa l'ozione telefonate vantaggiose verso l'estero
9	D_OP_NUM_PREF	Variabile che indica se è attiva, disattiva o dismessa l'ozione telefonate vantaggiose verso un numero preferito
10	D_RIC_RICORRENTE	Variabile che indica se è attiva, disattiva o dismessa l'ozione di ricariche ricorrente
11	ETA_CUSTOMER	Età del cliente
12	FLAG_OPZ_COUNTRY	Flag che indica se è stato scelto un particolare paese per effettuare chiamate vantaggiose
13	GENDER	Genere
14	ID_CUSTOMER	ID Cliente
15	MINUTI_ASSISTENZA	Minuti chiamate effettuate dal cliente per ricevere assistenza dall'operatore, negli ultimi 6 mesi
16	MINUTI_VOCE_ITZ	Minuti voce verso direttrici internazionali negli ultimi 6 mesi
17	MINUTI_VOCE_OFFNET	Minuti voce offnet (SIM di altri operatori) negli ultimi 6 mesi
18	MINUTI_VOCE_ONNET	Minuti voce onnet (SIM dello stesso operatore) negli ultimi 6 mesi
19	NUMERO_RICARCIHE_BONUS	Numero di ricariche bonus negli ultimi 6 mesi
20	NUMERO_RICARICHE_RICORRENTI	Numero di ricariche ricorrenti negli ultimi 6 mesi
21	RECENZA_RICARICA_BONUS	Mesi trascorsi dall'ultima volta in cui il cliente ha ricevuto una ricarica bonus
22	REC_CONT_INBOUND	Mesi trascorsi dall'ultima volta in cui il cliente ha contattato il call center
23	REC_CONT_OUTBOUND	Mesi trascorsi dall'ultima volta in cui il call center ha contattato il cliente
24	SIM_ATTIVE	Numero di SIM attive per cliente
25	ARPU	Valore arpu: ricavi medi ottenuti mensilmente per ciascun utente

Esercizio

1. Cambiare la directory di lavoro in quella in cui si è salvato il dataset e salvare il dataset in un oggetto in R.
2. Effettuare un'analisi fattoriale utilizzando le seguenti variabili:

CONTATTI_INBOUND
CONTATTI_OUTBOUND
REC_CONT_INBOUND
REC_CONT_OUTBOUND
MINUTI_ASSISTENZA
MINUTI_VOCE_ITZ
MINUTI_VOCE_OFFNET
MINUTI_VOCE_ONNET
RECENZA_RICARICA_BONUS
AMMONT_RICARICA_BONUS
AMMONT_RICARICA_PAG
AMMONT_RICARICA_PAG_LOTTO_SISAL
AMMONT_RICARICA_RICORRENTE
NUMERO_RICARCIHE_BONUS
NUMERO_RICARICHE_RICORRENTI
FLAG_OPZ_COUNTRY

Esercizio

→ Scegliere il numero di fattori ottimali

→ Salvare i fattori interpretati in un nuovo dataset

3. Stimare un modello di regressione lineare utilizzando

→come variabile dipendente il valore dell'Arpu

→ come potenziali regressori, oltre ai fattori individuati al punto precedente, anche le variabili: età del cliente, anzianità della sim e numero di sim attive per cliente:

- Utilizzare l'opzione di stepwise
- Effettuare tutti i passaggi presenti nelle slide di riepilogo, rispondendo anche alle seguenti domande:

    a. Il valore dell'R-quadro è soddisfacente?

    b. Cosa possiamo affermare osservando i dati relativi al test F e al test t?

    c. Quale regressore influenza maggiormente la variabile dipendente?