

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:

Lunedì 11 Gennaio 2016

- 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;
 - CD-ROM o chiavetta USB contenente:
 - questionario;
 - base dati in formato Excel;
 - programma SAS e output.

N.B. Il supporto elettronico (CD-ROM/chiavetta USB) non sarà restituito.

Consegna Lavoro di gruppo

- Si richiede di comunicare l'intenzione a svolgere l'homework facoltativo entro **Lunedì 14 Dicembre 2015**, via mail a gdeppieri@liuc.it e gmagistrelli@liuc.it.
- Il materiale utile a svolgere l'homework sarà consegnato in aula **Venerdì 18 Dicembre 2015**.
- La consegna dell'homework dovrà essere effettuata congiuntamente al lavoro di gruppo.

Processo di analisi

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



Selezione di k fattori
(dove $k < p$)

Utilizzo di alcuni criteri per la *selezione dei possibili di 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

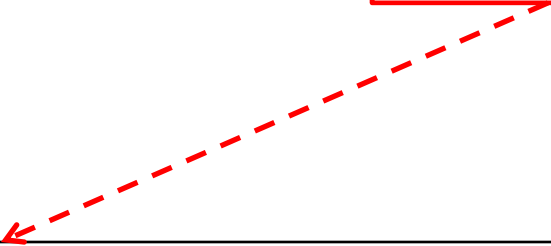
PROC FACTOR – Sintassi

Analisi fattoriale con il metodo delle componenti principali.

```
PROC FACTOR DATA=libreria.tabella option(s) ;
```

```
VAR elenco variabili ;
```

```
RUN ;
```



OPZIONE	DESCRIZIONE
PLOTS=SCREE(UNPACK)	Produce in output lo scree plot
FUZZ = <i>valore</i>	Nella matrice dei Loadings, stampa solo loadings > valore
N = <i>n</i>	Consente di specificare il numero di fattori che si vuole estrarre
OUT = <i>dataset</i>	Produce in output un dataset che contiene tutte le variabili originarie e i fattori non ruotati
ROTATE = <i>metodo</i>	Specifica il criterio da utilizzare per la rotazione dei fattori (es.VARIMAX)
REORDER	Nella matrice dei Loadings, ordina le variabili originarie in modo da facilitarne la lettura

PROC REG – Riepilogo

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. Analisi di influenza con i soli regressori scelti nella stepwise.
 - ✓ Se si è in presenza di osservazioni influenti: eliminarle e ripetere i punti 3 e 4
 - ✓ In assenza di osservazioni influenti: passare al punto 6

PROC REG – Riepilogo

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 $VIF_j = 1$)
 - ✓ Se si è in presenza di multicollinearità: azioni per eliminarla e ripetere i punti 3, 4, 5
 - ✓ 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)
8. Interpretazione dei coefficienti standardizzati

PROC REG – Sintassi

Modello di regressione lineare

```
proc reg data=dataset;  
    model variabile_dipendente=  
        regressore_1 ... regressore_p  
  
    /option(s);  
  
run;
```

OPTIONS:

- **STB** calcola i coefficienti standardizzati
- **selection=stepwise** applica la procedura stepwise per la selezione dei regressori
- **slentry=...** livello di significatività per testare l'entrata del singolo regressore nel modello
- **slstay=...** livello di significatività per testare la rimozione del singolo regressore dal modello
- **VIF** per verificare presenza di multicollinearità

PROC REG – Sintassi

La PROC REG fornisce nell'output i valori della **distanza di Cook** e del **leverage H** per ogni osservazione del dataset:

```
proc reg data=dataset noprint,  
  model variabile_dipendente=  
    regressore_1 ... regressore_p  
  / influence;  
output out=dataset_output cookd=cook H=leverage;  
run;
```

OPTIONS:

- **Influence** fornisce una serie di indicatori di influenza tra cui D e H
- **Cookd=** crea nel dataset di output una variabile con i valori della Distanza di Cook per ogni osservazione
- **H=** crea nel dataset di output una variabile con i valori del Leverage per ogni osservazione
- **Noprint** = utile soprattutto per dataset con molte informazioni, permette di non stampare l'output

Esercizio

Il dataset `ct_telefonia.sas7bdat` 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 contattato il call center negli ultimi 6 mesi
7	CONTATTI_OUTBOUND	Numero di volte in cui il call center ha contattato il cliente negli ultimi 6 mesi (per campagna commerciale)
8	D_OPZ_ESTERO	Variabile che indica se è attiva, disattiva o dismessa l'opzione telefonate vantaggiose verso l'estero
9	D_OP_NUM_PREF	Variabile che indica se è attiva, disattiva o dismessa l'opzione telefonate vantaggiose verso un numero preferito
10	D_RIC_RICORRENTE	Variabile che indica se è attiva, disattiva o dismessa l'opzione 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_RICARICHE_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. Allocare una libreria che punti alla cartella in cui si è salvato il dataset.
2. Accertarsi che le opzioni per l'output HTML siano correttamente impostate
3. 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_RICARICHE_BONUS
NUMERO_RICARICHE_RICORRENTI
FLAG_OPZ_COUNTRY

Esercizio

- Scegliere il numero di fattori ottimali
 - Salvare i fattori interpretati in un nuovo dataset
4. 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 (ed i relativi livelli di significatività)
 - 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 ai test t?
 - c. Quale regressore influenza maggiormente la variabile dipendente?