# PSP6075525 - Testing psicologico (matr. dispari)

## Esame del 300123

### Istruzioni

- Si avvii una nuova sessione di R (o RStudio).
- Si crei un nuovo script di R e lo si salvi come `cognome_nome.R`.
- Si effettui il download del file di dati dell'esame `dati_esame.Rdata` disponibile presso la pagina Moodle Esami del corso e lo si carichi nell'ambiente di lavoro di R.
- Si utilizzi il file `cognome_nome.R` per inserire il codice R utilizzato in risposta i quesiti d'esame. Attenzione: si ricorda di inserire il medesimo codice nel campo di risposta disponibile per ciascun quesito nel form Moodle dell'esame.
- Si invii il file `cognome_nome.R` mediante l'apposita funzione `Consegna codice R` presente nella pagina Moodle Esami del corso.
- Nota: la valutazione della prova sarà effettuata utilizzando primariamente il file `cognome_nome.R`. Si raccomanda pertanto la massima chiarezza nella scrittura delle risposte e la correttezza nel riportare i comandi e gli output di R per ciascun quesito d'esame.

Il file `data_exam.Rdata` contiene i dati relativi alla somministrazione del test AX001 ad un campione casuale di $n = 5000$ studenti frequentanti l'università di Teramo. Il test, somministrato per la valutazione delle abilità matematiche, è composto da $p = 15$ item rilevati su scale ordinali a 7 punti (livelli alti della scala indicano migliore performance matematica). Successivamente alla raccolta dei dati, le variabili osservate sono state adeguatamente quantificate mediante apposita procedura. L'obiettivo è quello di studiare la dimensionalità del test AX001 con particolare riferimento al numero e alla tipologia di dimensioni latenti che esso individua. Si importi il dataset in `R` e si risponda ai quesiti che seguono.

1. Si esegua una divisione a metà del dataset (50% di unità statistiche per la prima metà) e si utilizzi la prima metà del dataset per le analisi esplorative e la seconda metà per le analisi confermative. Successivamente si esegua un'analisi basata sul clustering gerarchico con metodo `ward.D2` e si individuino un numero congruo di raggruppamenti delle variabili osservate. Sulla base dell'analisi di raggruppamento si proponga un modello CFA. Nota: si suggerisce di impostare il seed di generazione casuale pari a `seedx=16001`.

```
Y = psych::rescale(Y,0,1)
out = split_dataset(data = Y,prop = 0.50,seedx = 16001)
YA = out$A; YB = out$B
hc = hclust(d = dist(cor(YA)),method = "ward.D2")
#x11();plot(hc)
mod1_def = hclust2lavaan(tree = hc,ngroups = 2)
```

Il clustering gerarchico consente di individuare due raggruppamenti per le $p = 15$ variabili osservate. Il modello CFA corrispondente scritto secondo la sintassi di `lavaan` è il seguente: `eta1 =  Y1+Y2+Y5+Y7+Y8+Y15` `eta2 =  Y3+Y4+Y6+Y9+Y10+Y11+Y12+Y13+Y14`

2. Si adatti ai dati il modello CFA definito al punto precedente secondo la metrica ULI. Si commenti il risultato ottenuto anche alla luce dell'adattamento complessivo del modello ai dati. Nota: nell'interpretazione della soluzione fattoriale si utilizzino i coefficienti stimati standardizzati.

```
mod1_fit = cfa(model = mod1_def,data = YB)
summary_table(mod1_fit,standardized = TRUE)

            parameter    lhs op  rhs     est     se        z pvalue
   1  latent->manifest | eta1 =~   Y1  0.9307 0.0000
   2  latent->manifest | eta1 =~   Y2 -0.0097 0.0229  -0.4515 0.6517
   3  latent->manifest | eta1 =~   Y5  0.0189 0.0234   0.8769 0.3805
   4  latent->manifest | eta1 =~   Y7 -0.1415 0.0293  -5.2964      0
   5  latent->manifest | eta1 =~   Y8  0.3214 0.0454   7.5515      0
   6  latent->manifest | eta1 =~  Y15  0.1897 0.0327   6.2574      0
   7  latent->manifest | eta2 =~   Y3  0.7274 0.0000
   8  latent->manifest | eta2 =~   Y4  0.1113 0.0345   4.4673      0
   9  latent->manifest | eta2 =~   Y6  0.1442 0.0349   5.7321      0
   10 latent->manifest | eta2 =~   Y9  0.4068 0.0412  13.7974      0
   11 latent->manifest | eta2 =~  Y10  0.4350 0.0425  14.3613      0
   12 latent->manifest | eta2 =~  Y11  0.1767 0.0356   6.9412      0
   13 latent->manifest | eta2 =~  Y12  0.1487 0.0346   5.9004      0
   14 latent->manifest | eta2 =~  Y13  0.2105 0.0362    8.152      0
   15 latent->manifest | eta2 =~  Y14  0.2844 0.0378  10.5817      0
   16       covariance |   Y1 ~~   Y1  0.1338 0.1031   1.2929  0.196
   17       covariance |   Y2 ~~   Y2  0.9999 0.0277  35.3548      0
   18       covariance |   Y5 ~~   Y5  0.9996 0.0289  35.3534      0
   19       covariance |   Y7 ~~   Y7  0.9800 0.0289  35.1394      0
   20       covariance |   Y8 ~~   Y8  0.8967 0.0278  31.6748      0
   21       covariance |  Y15 ~~  Y15  0.9640 0.0278  34.7983      0
   22       covariance |   Y3 ~~   Y3  0.4710 0.0315  14.5048      0
   23       covariance |   Y4 ~~   Y4  0.9876 0.0277  35.1278      0
   24       covariance |   Y6 ~~   Y6  0.9792 0.0278  34.9687      0
```

```
   25      covariance |   Y9 ~~   Y9  0.8345 0.0267  31.4133      0
   26      covariance |  Y10 ~~  Y10  0.8108 0.0268  30.6285      0
   27      covariance |  Y11 ~~  Y11  0.9688 0.0280  34.7658      0
   28      covariance |  Y12 ~~  Y12  0.9779 0.0272  34.9437      0
   29      covariance |  Y13 ~~  Y13  0.9557 0.0280   34.501      0
   30      covariance |  Y14 ~~  Y14  0.9191 0.0277  33.7011      0
   31      covariance | eta1 ~~ eta1  1.0000 0.1067   8.0861      0
   32      covariance | eta2 ~~ eta2  1.0000 0.0376   13.663      0
   33      covariance | eta1 ~~ eta2 -0.4223 0.0196 -14.3801      0
```

```
fitMeasures(object = mod1_fit,fit.measures = c("CFI","RMSEA","df","npar"))
```

```
     cfi  rmsea     df   npar
   0.126  0.213 89.000 31.000
```

```
p = NCOL(YB)
100*(31/(p*(p+1)/2))
```

```
   [1] 25.83
```

L'adattamento del modello ai dati avviene mediante stimatori di massima verosmiglianza, essendo le variabili osservate già quantificate. Il modello adattato richiede 31 parametri da stimare mentre il numero di parametri totali ammonta a $p(p+1)/2 = 120$. I gradi di libertà del modello adattato sono $df = 31$ con un indice di parsimoniosità soddisfacente (il modello utilizza solo il $100 \times \frac{31}{120} = 26\%$ dei parametri totali). Le misure di adattamento globale ai dati, tuttavia, non sono adeguate (gli indici CFI e RMSEA sono fuori range). Ciò è naturalmente riflesso nella quantificazione dei due misurandi, non sempre adeguata rispetto ai legami strutturali ipotizzati (si noti la presenza di item con coefficienti fattoriali molto bassi). In maniera complementare, le varianze di errore $\texttt{diag}(\boldsymbol{\Theta}_\delta)$ sono di magnitudine elevata. Complessivamente il modello di misura, sebbene parsimonioso, necessita di una completa revisione, in paticolar modo per quanto concerne la definizione della struttura fattoriale.

3. Si semplifichi il modello CFA adattato al punto precedente utilizzando il criterio $\hat{\lambda}_{jk} < 0.17$. Si commenti il risultato ottenuto.

```
summary_table(mod1_fit,standardized = TRUE,type_summary = "latent")
```

```
            parameter    lhs op rhs     est     se       z pvalue
   1  latent->manifest | eta1 =~  Y1  0.9307 0.0000
   2  latent->manifest | eta1 =~  Y2 -0.0097 0.0229 -0.4515 0.6517
   3  latent->manifest | eta1 =~  Y5  0.0189 0.0234  0.8769 0.3805
   4  latent->manifest | eta1 =~  Y7 -0.1415 0.0293 -5.2964      0
   5  latent->manifest | eta1 =~  Y8  0.3214 0.0454  7.5515      0
   6  latent->manifest | eta1 =~ Y15  0.1897 0.0327  6.2574      0
   7  latent->manifest | eta2 =~  Y3  0.7274 0.0000
   8  latent->manifest | eta2 =~  Y4  0.1113 0.0345  4.4673      0
   9  latent->manifest | eta2 =~  Y6  0.1442 0.0349  5.7321      0
   10 latent->manifest | eta2 =~  Y9  0.4068 0.0412 13.7974      0
   11 latent->manifest | eta2 =~ Y10  0.4350 0.0425 14.3613      0
   12 latent->manifest | eta2 =~ Y11  0.1767 0.0356  6.9412      0
   13 latent->manifest | eta2 =~ Y12  0.1487 0.0346  5.9004      0
   14 latent->manifest | eta2 =~ Y13  0.2105 0.0362   8.152      0
   15 latent->manifest | eta2 =~ Y14  0.2844 0.0378 10.5817      0
```

```
mod2_def = "eta1 =~ Y1+Y8+Y15 \n eta2 =~ Y3+Y9+Y10+Y11+Y13+Y14"
mod2_fit = cfa(model = mod2_def,data = YB)
summary_table(mod2_fit,standardized = TRUE)
```

```
            parameter    lhs op  rhs     est     se       z pvalue
   1  latent->manifest | eta1 =~   Y1  1.0178 0.0000
   2  latent->manifest | eta1 =~   Y8  0.2975 0.0538  5.3958      0
   3  latent->manifest | eta1 =~  Y15  0.1743 0.0356  4.8322      0
   4  latent->manifest | eta2 =~   Y3  0.8475 0.0000
```

```
   5  latent->manifest | eta2 =~   Y9  0.3510 0.0347 12.1217      0
   6  latent->manifest | eta2 =~  Y10  0.4092 0.0375 13.1497      0
   7  latent->manifest | eta2 =~  Y11  0.2305 0.0307  9.0144      0
   8  latent->manifest | eta2 =~  Y13  0.1909 0.0298  7.7099      0
   9  latent->manifest | eta2 =~  Y14  0.2414 0.0312  9.3487      0
   10       covariance |   Y1 ~~   Y1 -0.0359 0.1792 -0.1994  0.842
   11       covariance |   Y8 ~~   Y8  0.9115 0.0295 30.3837      0
   12       covariance |  Y15 ~~  Y15  0.9696 0.0280 34.7396      0
   13       covariance |   Y3 ~~   Y3  0.2817 0.0437  6.2584      0
   14       covariance |   Y9 ~~   Y9  0.8768 0.0266 33.0561      0
   15       covariance |  Y10 ~~  Y10  0.8326 0.0266 31.6315      0
   16       covariance |  Y11 ~~  Y11  0.9468 0.0275 34.6114      0
   17       covariance |  Y13 ~~  Y13  0.9636 0.0279 34.8797      0
   18       covariance |  Y14 ~~  Y14  0.9417 0.0278 34.5219      0
   19       covariance | eta1 ~~ eta1  1.0000 0.1814  5.6891      0
   20       covariance | eta2 ~~ eta2  1.0000 0.0505 13.8256      0
   21       covariance | eta1 ~~ eta2 -0.3112 0.0201 -13.127      0
```

Il modello semplificato e adattato ai dati presenta problemi di convergenza come evidenziato dal fatto che $\hat{\delta}_{11} = -0.0359$.

4. Si consideri l'insieme delle variabili osservate utilizzate al punto precedente. Si definisca un modello CFA con $q = 1$ variabili latenti e lo si adatti ai dati secondo la metrica ULI. Successivamente si confronti il risultato ottenuto con quello del punto precedente e si scelga quale dei due modelli è da preferire.

```
mod3_def = "eta =~ Y1+Y8+Y15+Y3+Y9+Y10+Y11+Y13+Y14"
mod3_fit = cfa(model = mod3_def,data = YB,check.post=TRUE)
summary_table(mod3_fit)

             parameter   lhs op rhs     est     se        z pvalue
   1  latent->manifest | eta =~  Y1  0.4998 0.0000
   2  latent->manifest | eta =~  Y8  0.3448 0.0587   11.673      0
   3  latent->manifest | eta =~ Y15  0.2286 0.0545   8.4261      0
   4  latent->manifest | eta =~  Y3 -0.5482 0.0714 -15.1695      0
   5  latent->manifest | eta =~  Y9 -0.5516 0.0729 -15.1999      0
   6  latent->manifest | eta =~ Y10 -0.2368 0.0550  -8.6829      0
   7  latent->manifest | eta =~ Y11 -0.0575 0.0508  -2.2764 0.0228
   8  latent->manifest | eta =~ Y13 -0.3367 0.0591 -11.4758      0
   9  latent->manifest | eta =~ Y14 -0.3502 0.0600 -11.8023      0
   10       covariance |  Y1 ~~  Y1  0.7502 0.0264  28.3313      0
   11       covariance |  Y8 ~~  Y8  0.8811 0.0266  32.5532      0
   12       covariance | Y15 ~~ Y15  0.9477 0.0278  34.2175      0
   13       covariance |  Y3 ~~  Y3  0.6995 0.0259  26.2625      0
   14       covariance |  Y9 ~~  Y9  0.6957 0.0267  26.1015      0
   15       covariance | Y10 ~~ Y10  0.9439 0.0280  34.1292      0
   16       covariance | Y11 ~~ Y11  0.9967 0.0284   35.287      0
   17       covariance | Y13 ~~ Y13  0.8866 0.0274  32.7008      0
   18       covariance | Y14 ~~ Y14  0.8774 0.0275  32.4515      0
   19       covariance | eta ~~ eta  1.0000 0.0244  10.1931      0

fitMeasures(object = mod3_fit,fit.measures = c("CFI","RMSEA","df","npar"))

      cfi  rmsea     df   npar
    0.374  0.175 27.000 18.000

p = NCOL(YB)
100*(18/(p*(p+1)/2))

   [1] 15
```

Il modello unidimensionale è naturalmente più parsimonioso del precedente, utilizzando solo il $100 \times \frac{18}{120} = 15\%$ dei parametri totali. L'adattamento ai dati del modello unidimensionale, tuttavia, non è ancora soddisfacente come evidenziato dagli indici RMSEA e CFI. Rispetto al modello definito al punto precedente, il modello unidimensionale è da preferire, essendo il primo affetto da problemi di convergenza.

5. Si consideri il modello unidimensionale definito al punto precedente e lo si migliori aggiungendo come parametri da stimare le covarianze di errore a coppia (ad esempio, sull'insieme $\{Y_1, Y_2, Y_3, Y_4\}$ si considerino <u>solo</u> le quantità COV$(Y_1, Y_2)$,COV$(Y_2, Y_3)$, COV$(Y_3, Y_4)$). Si confronti il nuovo modello con quello unidimensionale del punto precedente rispetto (i) al fit complessivo e (ii) all'errore di previsione tramite metodo Monte Carlo. Si scelga, con adeguata giustificazione, il modello finale. Nota: per il calcolo dell'errore di previsione si utilizzi `k=7` e `B=250`.

```r
mod4_def = "eta =~ Y1+Y8+Y15+Y3+Y9+Y10+Y11+Y13+Y14
Y1~~Y8
Y8~~Y15
Y15~~Y3
Y3~~Y9
Y9~~Y10
Y10~~Y11
Y11~~Y13
Y13~~Y14"
mod4_fit = cfa(model = mod4_def,data = YB,check.post=TRUE)
summary_table(mod4_fit)
```

```
            parameter   lhs op rhs     est     se        z pvalue
   1  latent->manifest | eta =~  Y1  0.4363 0.0000
   2  latent->manifest | eta =~  Y8  0.3641 0.0597  13.9347      0
   3  latent->manifest | eta =~ Y15  0.1847 0.0582   7.2785      0
   4  latent->manifest | eta =~  Y3 -0.3888 0.0729 -12.0896      0
   5  latent->manifest | eta =~  Y9 -0.6557 0.0963 -15.8291      0
   6  latent->manifest | eta =~ Y10 -0.1720 0.0623  -6.3827      0
   7  latent->manifest | eta =~ Y11  0.0370 0.0597   1.4339 0.1516
   8  latent->manifest | eta =~ Y13 -0.4263 0.0769 -12.8324      0
   9  latent->manifest | eta =~ Y14 -0.4766 0.0805 -13.7185      0
   10       covariance |  Y1 ~~  Y8  0.1890 0.0186   8.4323      0
   11       covariance |  Y8 ~~ Y15 -0.1999 0.0184  -9.8763      0
   12       covariance | Y15 ~~  Y3 -0.1515 0.0181  -7.4649      0
   13       covariance |  Y3 ~~  Y9  0.1670 0.0200   5.8162      0
   14       covariance |  Y9 ~~ Y10 -0.3540 0.0186  -14.461      0
   15       covariance | Y10 ~~ Y11  0.1570 0.0202   7.7376      0
   16       covariance | Y11 ~~ Y13  0.1210 0.0194   5.7288      0
   17       covariance | Y13 ~~ Y14 -0.4847 0.0205 -19.0994      0
   18       covariance |  Y1 ~~  Y1  0.8097 0.0254  31.7298      0
   19       covariance |  Y8 ~~  Y8  0.8675 0.0260  32.9712      0
   20       covariance | Y15 ~~ Y15  0.9659 0.0276  34.8487      0
   21       covariance |  Y3 ~~  Y3  0.8488 0.0265  31.1961      0
   22       covariance |  Y9 ~~  Y9  0.5700 0.0284  20.5444      0
   23       covariance | Y10 ~~ Y10  0.9704 0.0284  34.6847      0
   24       covariance | Y11 ~~ Y11  0.9986 0.0286  35.3282      0
   25       covariance | Y13 ~~ Y13  0.8183 0.0276  30.1061      0
   26       covariance | Y14 ~~ Y14  0.7729 0.0275  28.5799      0
   27       covariance | eta ~~ eta  1.0000 0.0200   9.4901      0
```

```r
fitMeasures(object = mod4_fit,fit.measures = c("CFI","RMSEA","df","npar"))
```

```
     cfi  rmsea     df   npar
   0.668  0.152 19.000 26.000
```

```r
p = NCOL(YB)
100*(26/(p*(p+1)/2))
```

```
  [1] 21.67
```

```r
fitMeasures(object = mod3_fit,fit.measures = c("CFI","RMSEA","df","npar"))
```

```
     cfi  rmsea     df   npar
   0.374  0.175 27.000 18.000
```

```
fitMeasures(object = mod4_fit,fit.measures = c("CFI","RMSEA","df","npar"))

      cfi  rmsea     df   npar
    0.668  0.152 19.000 26.000

YB_ridotto = YB[,c("Y1","Y8","Y15","Y3","Y9","Y10","Y11","Y13","Y14")]
kf1 = kFold_validation(model_definition = mod3_def,
                       data = YB_ridotto,
                       nfold = 7,error = "montecarlo",
                       B = 250,force_crossValid = TRUE)

Note: Computing prediction error using montecarlo approach.
Done. Number of failures to convergence for training model: 0/7

kf2 = kFold_validation(model_definition = mod4_def,
                       data = YB_ridotto,
                       nfold = 7,error = "montecarlo",
                       B = 250,force_crossValid = TRUE)

Note: Computing prediction error using montecarlo approach.
Done. Number of failures to convergence for training model: 0/7

summary(cbind(kf1,kf2))

         kf1             kf2
    Min.   :1.03   Min.   :0.939
    1st Qu.:1.50   1st Qu.:1.354
    Median :1.85   Median :1.501
    Mean   :1.77   Mean   :1.596
    3rd Qu.:2.07   3rd Qu.:1.959
    Max.   :2.39   Max.   :2.103

apply(cbind(kf1,kf2),2,sd) / apply(cbind(kf1,kf2),2,mean)

      kf1    kf2
   0.2607 0.2677
```

Il nuovo modello prevede l'aggiunta di 15 parametri aggiuntivi ed è meno parsimonioso di quello unidimensionale (21% di parametri stimati contro il 15% precedente). Come atteso, il confronto in termini di adattamento complessivo è a favore del modello corrente sebbene l'indice RMSEA sia lievemente più basso di quello del modello unidimensionale. In termini di errore di previsione, i due modelli presentano variabilità pressocché analoghe (come evidenziato dai coefficienti di variazione) sebbene il modello corrente presenti indici di tendenza centrale e quartili più bassi del modello unidimensionale. La scelta finale dunque ricade sul modello corrente che necessiterebbe comunque di miglioramenti strutturali prima di un suo utilizzo pratico.