PSP6075525 - Testing psicologico (matr. dispari)

Prova d'esame del 13-11-20

Istruzioni iniziali

- 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 del corso e lo si carichi nell'ambiente di lavoro di R.
- Si crei un nuovo documento di testo (mediante LibreOffice Writer, Microsoft Word o software analogo) e lo si salvi come `cognome_nome.doc`. Il file dovrà contenere le risposte ai quesiti d'esame accompagnati dai comandi di R, dai risultati ottenuti e dai grafici prodotti. Le risposte dovranno essere inserite in ordine, rispettando il numero del quesito a cui si riferiscono. Alla fine, il file dovrà essere convertito in formato non modificabile (PDF: `cognome_nome.pdf`) ed inviato al docente utilizzando la procedura "Consegna documento" disponibile presso la pagina moodle del corso.
- La valutazione della prova sarà effettuata utilizzando primariamente il file `cognome_nome.pdf`: si raccomanda pertanto la chiarezza nella scrittura delle risposte e la correttezza nel riportare i comandi e gli output di R. Il file `cognome_nome.R` dovrà essere allegato al file `cognome_nome.pdf` solo per un controllo aggiuntivo (pertanto non verrà valutato primariamente).

Traccia d'esame

Il dataset contiene 7 variabili riferite ad un test di abilità cognitivo denominato *cog19test* somministrato ad un campione di $n = 250$ alunni di una scuola elementare della provincia di Padova. Le variabili originarie sono state pre-trattate mediante una adeguata procedura di quantificazione. L'obiettivo dell'analisi è quello di definire e adattare un modello fattoriale confermativo per lo studio della dimensionalità del test *cog19test*, con il particolare interesse a capire se il test sia unidimensionale o meno.

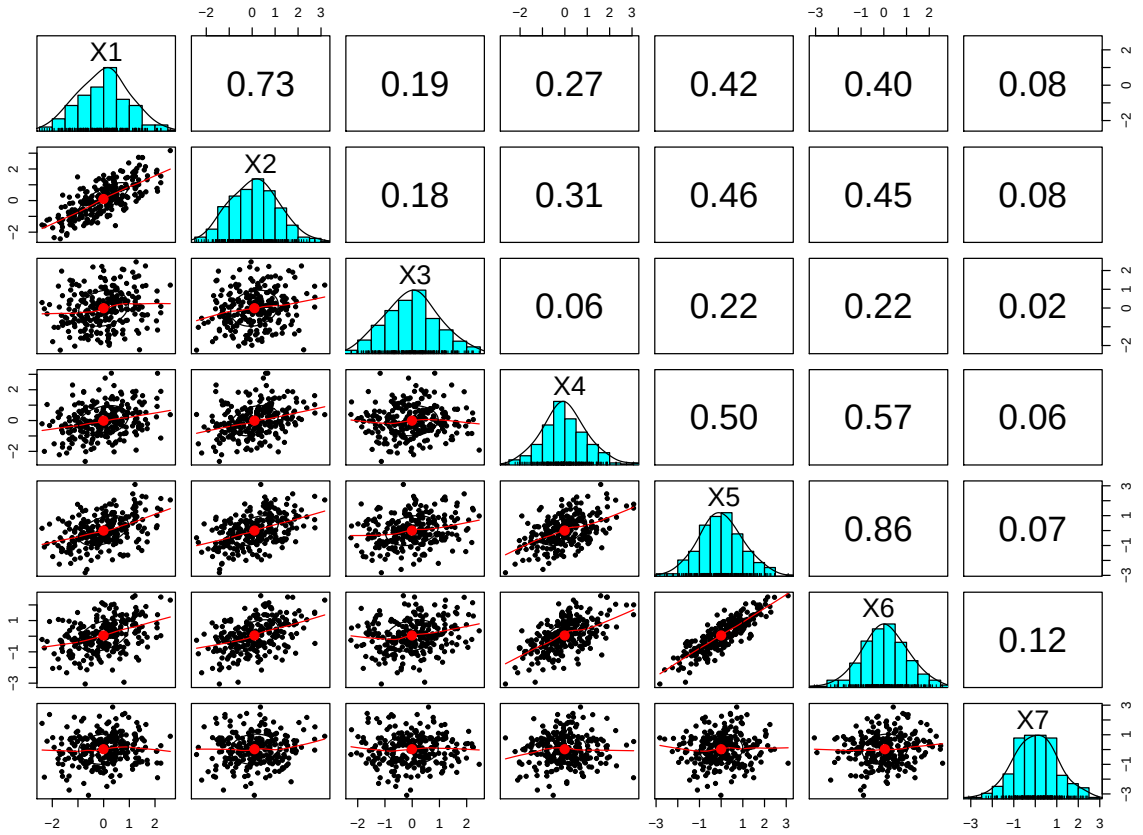
1. Si individuino il numero di unità statistiche e di variabili a disposizione, indicando per queste ultime il tipo di variabili coinvolte.
   Il numero di unità statistiche è pari a $n = 250$ mentre le variabili coinvolte sono $p = 7$, tutte di tipo numerico (variabili reali).

2. Si calcoli un'opportuna statistica di sintesi della matrice dei dati e si proponga una sua rappresentazione grafica.
   Una statistica opportuna di sintesi della matrice dei dati $\mathbf{X}$ è la covarianza $\boldsymbol{\Sigma}_{p\times p}$ o correlazione $\mathbf{R}_{p\times p}$. La prima può essere rappresentata mediante un grafico di dispersione a coppie mentre la seconda mediante un grafico tipo `corplot`. Ad esempio, considerando la matrice di covarianza possiamo calcolarla e rappresentarla graficamente come segue:

```
cov(X)
```

```
             X1         X2         X3         X4         X5         X6         X7
X1 0.93448352 0.73471468 0.17676453 0.25862758 0.41098238 0.3854728 0.08114452
X2 0.73471468 1.07819399 0.18427666 0.31098171 0.47799725 0.4699931 0.08101741
X3 0.17676453 0.18427666 0.93002779 0.05822807 0.21609570 0.2108494 0.01860730
X4 0.25862758 0.31098171 0.05822807 0.96377259 0.49229321 0.5654492 0.06397807
X5 0.41098238 0.47799725 0.21609570 0.49229321 1.00150022 0.8613202 0.07285333
X6 0.38547281 0.46999313 0.21084938 0.56544919 0.86132024 1.0114101 0.11834523
X7 0.08114452 0.08101741 0.01860730 0.06397807 0.07285333 0.1183452 1.04151074
```

```
psych::pairs.panels(X)
```

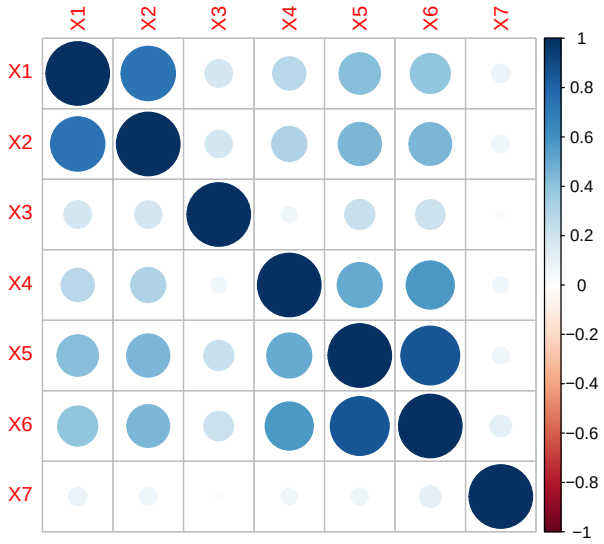


La matrice di correlazione invece è ottenibile e rappresentabile come segue:

```
cor(X)
```

```
             X1         X2         X3         X4         X5         X6         X7
X1 1.00000000 0.73195509 0.18961004 0.27252210 0.42482675 0.3965010 0.08225106
X2 0.73195509 1.00000000 0.18402377 0.30506972 0.45999321 0.4500693 0.07645362
X3 0.18961004 0.18402377 1.00000000 0.06150312 0.22390956 0.2174006 0.01890617
X4 0.27252210 0.30506972 0.06150312 1.00000000 0.50108458 0.5727204 0.06385751
X5 0.42482675 0.45999321 0.22390956 0.50108458 1.00000000 0.8558063 0.07133325
X6 0.39650099 0.45006932 0.21740059 0.57272039 0.85580630 1.0000000 0.11530688
X7 0.08225106 0.07645362 0.01890617 0.06385751 0.07133325 0.1153069 1.00000000
```

```
corrplot::corrplot(cor(X), method = "circle")
```



3. Si definisca un modello CFA ad una sola variabile latente, lo si adatti ai dati a disposizione e se ne valuti l'adattamento con almeno un indice opportuno.
   Il modello CFA ad una sola variabile latente è definito dall'equazione lineare

$$\mathbf{x}_i = \boldsymbol{\eta}_i \boldsymbol{\Lambda} + \boldsymbol{\delta}_i$$

   quando le osservazioni $i = 1, \ldots, n$ sono indipendenti e distribuite identicamente. Notiamo come la matrice $\boldsymbol{\Lambda}$ abbia dimensione $q \times p$ mentre il vettore $\boldsymbol{\eta}_i$ delle variabili aleatorie latenti sia di dimensione $1 \times q$. L'adattamento del modello ai dati è effettuato mediante la libreria `lavaan` come segue:

```
mod1_def = "eta =~ X1+X2+X3+X4+X5+X6+X7"
mod1_fit = lavaan::cfa(model = mod1_def,data = X)
```

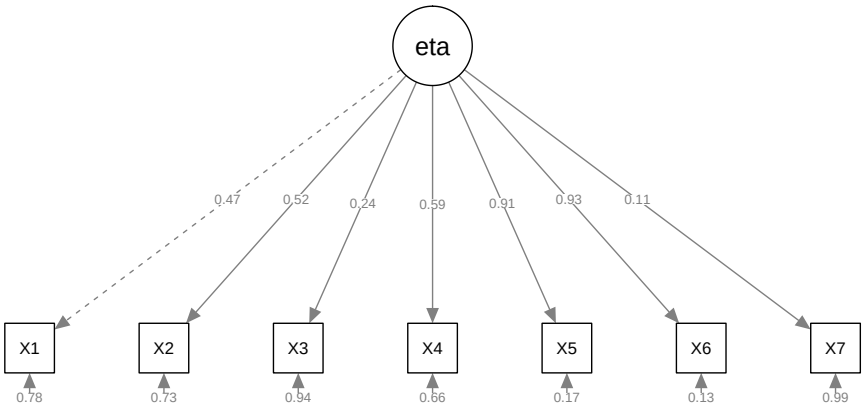
   Dopo aver adattato il modello ai dati a disposizione, vale a dire dopo aver stimato i parametri del modello $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}_\delta$, l'adattamento globale può essere valutato ad esempio mediante gli indici RMSEA o CFI:

```
lavaan::fitmeasures(object = mod1_fit,fit.measures = c("cfi","rmsea","df"))

     cfi  rmsea     df
   0.801  0.199 14.000
```

   Il valore dell'indice RMSEA basato sui residui del modello adattato indica che quest'ultimo non è globalmente soddisfacente.

4. Si rappresenti graficamente il modello adattato al punto 3 e si fornisca un'interpretazione dei risultati utilizzando la matrice $\boldsymbol{\Lambda}$ e $\boldsymbol{\Theta}_\delta$.

```
semPlot::semPaths(object = mod1_fit, whatLabels = "std.all",style = "lisrel")
```



I parametri stimati del modello possono essere estratti da un oggetto della classe `lavaan` mediante la funzione `inspect()`. Possono essere richiesti i valori stimati grezzi (..`what='est'`) oppure i valori stimati standardizzati (..`what='std.all'`). L'output della funzione `inspect()` è una lista.

```
out = lavaan::inspect(object = mod1_fit,what="std.all")
out$lambda #matrice Lambda

       eta
   X1 0.473
   X2 0.518
   X3 0.240
   X4 0.586
   X5 0.910
   X6 0.935
   X7 0.108

out$theta #matrice Theta_delta

      X1    X2    X3    X4    X5    X6    X7
   X1 0.776
   X2 0.000 0.731
   X3 0.000 0.000 0.943
   X4 0.000 0.000 0.000 0.657
   X5 0.000 0.000 0.000 0.000 0.172
   X6 0.000 0.000 0.000 0.000 0.000 0.127
   X7 0.000 0.000 0.000 0.000 0.000 0.000 0.988

out$psi #matrice Phi

      eta
   eta 1
```

Il modello CFA adattato è un modello con $q = 1$ variabili latenti e $p = 7$ variabili osservate/manifeste. In generale, il modello adattato non presenta indici di adattamento globale soddisfacente (vedi punto 3). La matrice $\boldsymbol{\Lambda}$ in questo caso specifico ha dimensione $p \times 1$ e contiene i coefficienti fattoriali del modello. Si ricordi che le scale dei parametri sono standardizzate nell'intervallo $[0, 1]$. Notiamo come, ad eccezione di alcune variabili (`X5`,`X6`), le variabili presentano coefficienti fattoriali molto bassi, alcuni prossimi allo zero (`X7`). Allo stesso modo, ciò si riflette sulla matrice delle varianze degli errori che presenta valori

verso 1, ad eccezione di `X5` e `X6`. La matrice delle varianze-covarianze in questo caso specifico perde di significato poiché il modello suppone $q = 1$ variabili latenti. Globalmente i risultati suggeriscono che il test in oggetto di valutazione non può dirsi unidimensionale.

5. Si adatti un modello CFA a due variabili latenti secondo l'assegnazione $\eta_1 : X1, X2, X3, X4$ e $\eta_2 : X5, X6, X7$. Si valuti l'adattamento del nuovo modello.

```
mod2_def = "eta1 =~ X1+X2+X3+X4 \n eta2 =~ X5+X6+X7"
mod2_fit = lavaan::cfa(model = mod2_def,data = X)
lavaan::fitmeasures(object = mod2_fit,fit.measures = c("cfi","rmsea","df"))
```

```
    cfi  rmsea     df
  0.909  0.139 13.000
```

Il nuovo modello, che rispetto al precedente è composto da $q = 2$ variabili latenti, presenta ancora un adattamento globale insoddisfacente rispetto ai valori degli indici RMSEA e CFI.

6. Si confronti opportunamente il modello adattato al punto 5 rispetto al modello adattato al punto 3. Quale dei due modelli risulta più parsimonioso e quale da preferire?

   I due modelli possono essere confrontati in termini di adattamento globale ai dati, ad esempio mediante l'uso ad esempio dell'indice RMSEA o AIC.

```
fm1 = lavaan::fitmeasures(object=mod1_fit,fit.measures=c("cfi","rmsea","df","AIC"))
fm2 = lavaan::fitmeasures(object=mod2_fit,fit.measures=c("cfi","rmsea","df","AIC"))
fm = rbind(fm1,fm2); rownames(fm) = c("mod1","mod2")
print(fm)
```

```
             cfi     rmsea df      aic
mod1 0.8005546 0.1986830 14 4413.601
mod2 0.9090821 0.1392085 13 4339.421
```

Notiamo come il secondo modello presenti indici di adattamento globali migliori del primo modello anche se ancora lontani dall'essere soddisfacenti. Il valore dell'indice AIC suggerisce come il secondo modello sia da preferire per successive modifiche. I gradi di libertà dei due modelli (`df`) differiscono poiché il primo modello ha meno parametri da stimare (liberi) rispetto al secondo modello.

7. Si utilizzi una procedura razionale per migliorare il modello adattato al punto 5 e si individui il modello che meglio si adatta ai dati rispetto a quest'ultimo.

   Una procedura razionale per migliorare il modello, quando non si dispongono di informazioni qualitative sulla struttura dimensionale di un test, è quella che prevede l'utilizzo dei c.d. indici di modifica. Un indice di modifica è il risultato di un test fatto sulla struttura fattoriale corrente rispetto alle strutture fattoriali che si otterrebbero se venissero stimati dei parametri assenti nella forma attuale. La statistica test utilizzata segue in distribuzione la *t-Student*: valori della statistica $|T| > 4$ possono essere utilizzati per aggiungere il parametro corrispondente al test effettuato. La procedura è implementata dalla funzione `modificationindices()` della libreria `lavaan`.

```
head(modificationindices(object = mod2_fit,sort. = TRUE),n = 10)
```

```
      lhs op rhs     mi    epc sepc.lv sepc.all sepc.nox
24 eta2 =~  X4 54.372  0.635   0.591    0.603    0.603
25   X1 ~~  X2 50.291  0.987   0.987    3.528    3.528
```

```
   41   X4 ~~  X6 21.147  0.137   0.137    0.384    0.384
   21 eta2 =~  X1  7.670 -0.284  -0.264   -0.274   -0.274
   32   X2 ~~  X4  5.262 -0.114  -0.114   -0.248   -0.248
   27   X1 ~~  X4  4.989 -0.100  -0.100   -0.202   -0.202
   22 eta2 =~  X2  4.825 -0.260  -0.242   -0.233   -0.233
   23 eta2 =~  X3  4.089  0.178   0.166    0.172    0.172
   29   X1 ~~  X6  2.976 -0.040  -0.040   -0.184   -0.184
   44   X5 ~~  X7  2.107 -0.050  -0.050   -0.135   -0.135
```

Notiamo dalla colonna `mi` che il parametro da aggiungere è $\lambda_{4,2}$ relativo al legame tra la variabile latente η_2 e la manifesta X_4 (colonne: `lhs`, `op`, `rhs`). La procedura suggerisce anche altre modifiche da apportare, ad esempio aggiungere il parametro $\theta_{\delta_{1,2}}$ relativo alla correlazione tra gli errori delle variabili manifeste X_1 e X_2. Procediamo, per il momento, aggiungendo un parametro alla volta.

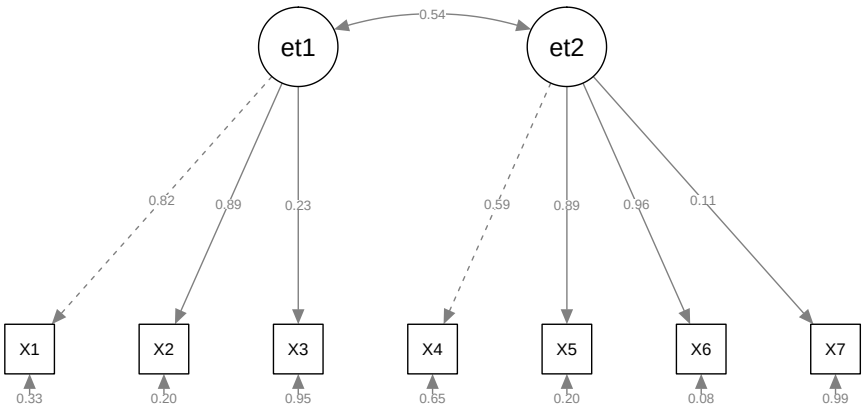
```
mod3_def = "eta1=~X1+X2+X3 \n eta2=~X4+X5+X6+X7"
mod3_fit = lavaan::cfa(model = mod3_def,data = X)

fm1 = lavaan::fitmeasures(object=mod2_fit,fit.measures=c("cfi","rmsea","df","AIC"))
fm2 = lavaan::fitmeasures(object=mod3_fit,fit.measures=c("cfi","rmsea","df","AIC"))
fm = rbind(fm1,fm2); rownames(fm) = c("mod1","mod2")
print(fm)
```

```
              cfi      rmsea df      aic
   mod1 0.9090821 0.13920849 13 4339.421
   mod2 0.9982072 0.01954837 13 4277.681
```

Il nuovo modello presenta valori di adattamento globale decisamente migliori rispetto al modello adattato al punto 5. Poiché la struttura fattoriale risultante è semplice nel senso dell'interpretabilità dei risultati decidiamo di non procedere oltre con il miglioramento del modello corrente. Si noti come la decisione di aggiungere il legame $\eta_2 : X_4$ è basata sul risultato di un test statistico: occorre sempre valutare se tale miglioramento sia sensato nel senso delle ipotesi sulla natura del test che si sta valutando.

8. Si rappresenti graficamente il modello finale scelto al punto 7 e lo si interpreti.

```
semPlot::semPaths(object = mod3_fit, whatLabels = "std.all",style = "lisrel")
```



```
summary(mod3_fit,standardized=TRUE)

  lavaan 0.6-7 ended normally after 31 iterations

    Estimator                                         ML
    Optimization method                           NLMINB
    Number of free parameters                         15

    Number of observations                           250

  Model Test User Model:

    Test statistic                                14.242
    Degrees of freedom                                13
    P-value (Chi-square)                           0.357

  Parameter Estimates:

    Standard errors                             Standard
    Information                                 Expected
    Information saturated (h1) model          Structured

  Latent Variables:
                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    eta1 =~
      X1                1.000                               0.790    0.819
      X2                1.170    0.116   10.066    0.000    0.924    0.892
      X3                0.277    0.082    3.367    0.001    0.219    0.228
    eta2 =~
      X4                1.000                               0.577    0.589
      X5                1.547    0.149   10.406    0.000    0.893    0.894
      X6                1.665    0.160   10.384    0.000    0.961    0.957
      X7                0.196    0.116    1.690    0.091    0.113    0.111

  Covariances:
                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    eta1 ~~
      eta2              0.247    0.045    5.479    0.000    0.541    0.541

  Variances:
                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
     .X1                0.307    0.060    5.092    0.000    0.307    0.330
     .X2                0.220    0.076    2.892    0.004    0.220    0.204
     .X3                0.878    0.079   11.083    0.000    0.878    0.948
     .X4                0.627    0.058   10.742    0.000    0.627    0.653
     .X5                0.201    0.037    5.442    0.000    0.201    0.201
     .X6                0.084    0.038    2.219    0.027    0.084    0.084
     .X7                1.025    0.092   11.172    0.000    1.025    0.988
      eta1              0.624    0.095    6.558    0.000    1.000    1.000
      eta2              0.333    0.067    4.971    0.000    1.000    1.000
```

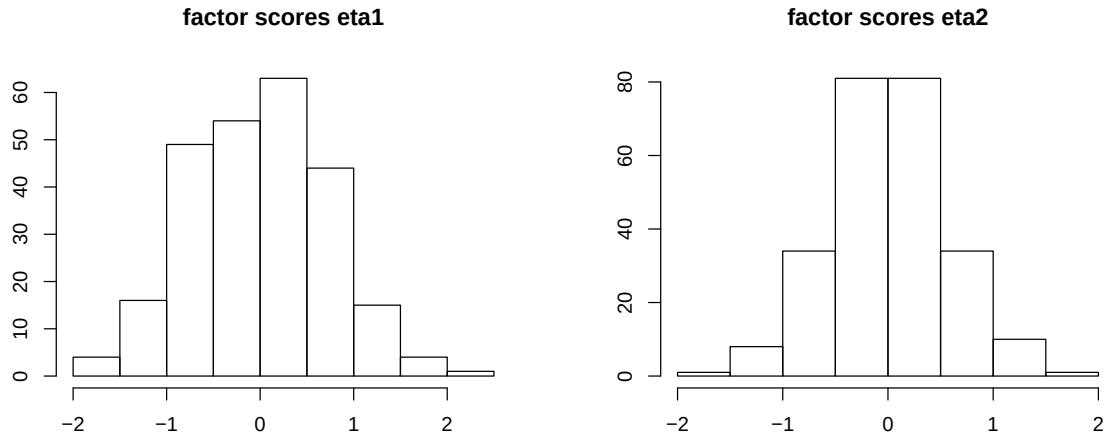
Il modello finale è composto da $q = 2$ variabili latenti e $p = 7$ variabili manifeste. L'adattamento globale è soddisfancente (RMSEA=0.02) come anche la parsimoniosità (parametri stimati = 13, parametri complessivi struttura satura = $p(p+1)/2 = 28$). Interpretanto la matrice dei coefficienti $\boldsymbol{\Lambda}$ notiamo come le

variabili manifeste presentano coefficienti fattoriali (standardizzati) adeguati, ad eccezione della variabile X_7. Le variabili latenti presentano una correlazione pari a $\phi_{2,1} = 0.541$. Si noti come alcune varianze dei residui del modello risultano ancora elevate (es.: $\theta_{\delta_{3,3}}$ e $\theta_{\delta_{7,7}}$).

9. Si calcolino i valori predetti dal modello finale a livello delle unità statistiche (c.d. *factor scores*) e li si rappresenti graficamente. Cosa possiamo dire circa la loro forma distributiva? Si fornisca una spiegazione basata sul razionale del modello CFA.

   I valori latenti predetti dal modello $\hat{\eta}_i = \mathbb{E}[\boldsymbol{\eta}_i|\mathbf{x}_i]$ possono essere calcolati in diversi modi. Un modo è quello di usare uno stimatore lineare basato su una procedura di regressione. In `lavaan` tale metodo è implementato mediante `lavPredict(..,method='regression')`.

```
Xpred = lavPredict(object = mod3_fit,newdata = X,type = "lv",method = "regression")
par(mfrow=c(1,2))
hist(Xpred[,1],main="factor scores eta1",ylab="",xlab="")
hist(Xpred[,2],main="factor scores eta2",ylab="",xlab="")
```




```
summary(Xpred)
```

```
      eta1                eta2          
 Min.   :-1.752212   Min.   :-1.7567  
 1st Qu.:-0.549736   1st Qu.:-0.3430  
 Median : 0.008716   Median : 0.0029  
 Mean   : 0.000000   Mean   : 0.0000  
 3rd Qu.: 0.510715   3rd Qu.: 0.3574  
 Max.   : 2.308140   Max.   : 1.5289  
```

   I punteggi fattoriali presentano distribuzione simmetrica e centrata sullo zero. Ciò dipende dal modello CFA adattato ai dati: questo infatti non modella le medie dei fattori latenti ($\boldsymbol{\tau} = \mathbf{0}$).

10. Si calcoli mediante un opportuno indice l'attendibilità delle scale derivanti dal modello finale adattato e se ne interpreti il risultato.

```
semTools::reliability(mod3_fit)
```

```
              eta1      eta2
```

```
alpha  0.6402666 0.6929842
omega  0.7268192 0.7696287
omega2 0.7268192 0.7696287
omega3 0.7308298 0.7763055
avevar 0.5207442 0.5161601
```

Un indice idoneo per valutare l'attendibilità delle scale η_1 ed η_2 secondo il principio della coerenza interna è l'indice $\omega \in [0, 1]$. In questo caso, la funzione `reliability()` della libreria `semTools` restituisce in output diversi indici di attendibilità, tra cui diverse versioni (corrette) dell'indice ω (`omega`). I risultati suggeriscono che entrambe le scale del *cog19test* presentino buoni indici di attendibilità/precisione.