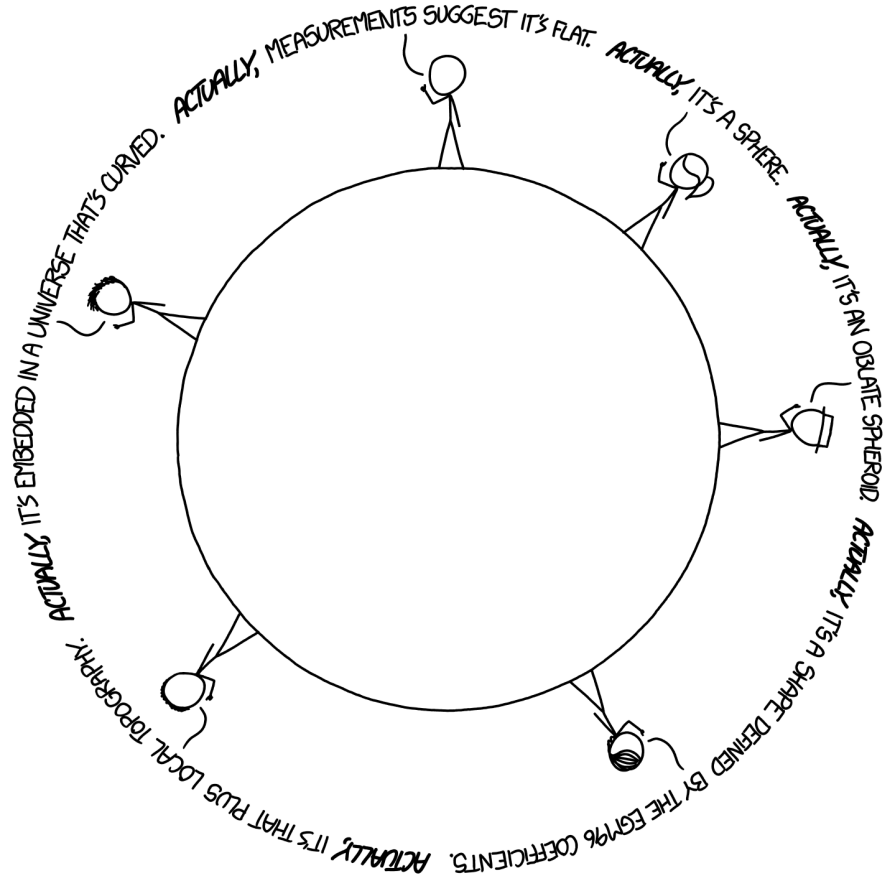
Corrado Caudek

Appunti di Costruzione e validazione di strumenti di misura dell'efficacia dell'intervento psicologico in neuropsicologia – B020881 (B213)

Appunti di Costruzione e validazione di strumenti di misura dell'efficacia dell'intervento psicologico in neuropsicologia – AA 2021/2022



Indice

Elenco delle figure

Elenco delle tabelle

Prefazione

La presente dispensa contiene il materiale delle lezioni dell'insegnamento di *Costruzione e validazione di strumenti di misura dell'efficacia dell'intervento psicologico in neuropsicologia* B020881 (B213) rivolto agli studenti del secondo anno del Corso di Laurea Magistrale in Psicologia Clinica e della Salute e Neuropsicologia (curriculum: assessment e intervento psicologici in neuropsicologia - E21), A.A. 2021-2022. L'insegnamento si propone di fornire agli studenti un'introduzione all'assessment psicologico, ovvero un insieme di conoscenze/competenze che si pongono all'intersezione tra psicometria, statistica e informatica.

Nello specifico, l'insegnamento si focalizzerà sull'analisi fattoriale confermativa (*confermatory factor analysis*, CFA) e sull'analisi fattoriale esplorativa, (*explorative factor analysis*, EFA), cioè sugli strumenti che vengono usati durante il processo di sviluppo dei test psicometrici, ovvero che vengono usati per esaminare la struttura latente di una scala psicologica (ad esempio un questionario). In questo contesto, la CFA viene utilizzata per verificare il numero di dimensioni sottostanti gli indicatori (fattori) e l'intensità delle relazioni item-fattore (saturazioni fattoriali). La CFA consente anche di capire di come dovrebbe essere svolto lo scoring di un test. Quando la struttura latente è multifattoriale (cioè, a due o più fattori), il numero di fattori è indicativo del numero di sottoscale e di come esse dovrebbero essere codificate. La CFA è un importante strumento analitico anche per altri aspetti della valutazione psicometrica. Può essere utilizzata per stimare l'affidabilità di scala dei test psicometrici in modo da evitare i problemi della teoria classica dei test (ad es. alpha di Cronbach). Dati i recenti progressi nell'analisi dei dati categoriali, ora la CFA offre un quadro analitico comparabile a quello offerto dalla teoria di risposta agli item (IRT). In effetti, secondo Brown (2015), la CFA offre una maggiore flessibilità analitica rispetto al modello IRT tradizionale.

Un costrutto è un concetto teorico che può essere operazionalizzato nei termini di un fattore. In psicologia clinica, psichiatria e neuropsicologia,

ad esempio, i disturbi mentali sono costrutti manifestati da vari insiemi di sintomi che sono riportati dal paziente o osservati da altri. La CFA è uno strumento analitico indispensabile per la validazione dei costrutti psicologici. I risultati della CFA possono fornire prove convincenti della validità convergente e discriminante dei costrutti teorici. La validità convergente è indicata dall'evidenza che diversi indicatori di costrutti teoricamente simili o sovrapposti sono fortemente correlati. La validità discriminante è indicata dai risultati che mostrano che gli indicatori di costrutti teoricamente distinti sono altamente incorrelati. Un punto di forza fondamentale degli approcci CFA per la costruzione e la validazione di uno strumento psicometrico è che le risultanti stime di validità convergente e discriminante sono corrette per l'errore di misurazione. Pertanto, la CFA fornisce un quadro analitico migliore rispetto ai metodi tradizionali che non tengono conto dell'errore di misurazione (ad esempio, gli approcci ordinari ai minimi quadrati come la correlazione/regressione multipla, i quali presuppongono che le variabili nell'analisi siano prive di errori di misurazione).

Spesso, parte della covariazione delle misure osservate è dovuta a fonti diverse dai fattori latenti di interesse. Questa covariazione aggiuntiva spesso riflette la varianza del metodo utilizzato per la misurazione. Gli effetti del metodo possono verificarsi anche all'interno di un'unica modalità di valutazione. Ad esempio, effetti del metodo sono solitamente presenti nei questionari che contengono una combinazione di elementi formulati positivamente e negativamente. Sfortunatamente, l'EFA non è in grado di stimare gli effetti del metodo. In effetti, l'uso di EFA quando esistono effetti del metodo può produrre risultati fuorvianti, ovvero suggerire la presenza di fattori aggiuntivi che corrispondono invece ad artefatti della misurazione. Nella CFA, invece, gli effetti del metodo possono essere specificati come parte della teoria dell'errore del modello di misurazione.

Un altro punto di forza della CFA è la sua capacità di affrontare il problema della generalizzabilità del modello di misurazione tra gruppi di individui o nel tempo. La valutazione dell'invarianza della misura è un aspetto importante dello sviluppo del test. Se un test è destinato a essere somministrato in una popolazione eterogenea, si dovrebbe stabilire che le sue proprietà di misurazione sono equivalenti in sottogruppi della popolazione (es. sesso, razza). Si dice che un test è distorto quando alcuni dei suoi elementi non misurano il costrutto sottostante in modo comparabile tra gruppi di rispondenti. Il test fornisce una stima distorta se,

ad esempio, per un dato livello di vera intelligenza, gli uomini tendono a ottenere un punteggio di QI più alto rispetto alle donne. Il problema della generalizzabilità della validità del costrutto tra i gruppi può essere affrontato nella CFA esaminando gruppi multipli mediante modelli MIMIC (indicatori multipli, cause multiple). Inotre, è possibile chiedersi se il modello di misurazione sia equivalente tra i gruppi. Le soluzioni CFA a gruppi multipli vengono anche utilizzate per esaminare l'invarianza della misurazione longitudinale. Questo è un aspetto molto importante dell'analisi delle variabili latenti dei progetti di misure ripetute. In assenza di tale valutazione, non è possibile determinare se il cambiamento temporale in un costrutto sia dovuto a un vero cambiamento dei rispondenti o a cambiamenti nel modo di rispondere alla scala nel tempo. L'analisi a gruppi multipli può essere applicata a qualsiasi tipo di modello CFA. Ad esempio, queste procedure possono essere incorporate nell'analisi dei dati multitratto-multimetodo per esaminare la generalizzabilità della validità del costrutto tra gruppi.

In questo insegnamento la discussione delle teciche della CFA sarà preceduta da un'introduzione relativa alla EFA e la teoria classica dei test. La EFA, infatti, può essere concepita il metodo che viene utilizzato nei primi passi dello sviluppo di una scala psicometria, mentre la teoria classica dei test rappresenta la cornice teorica di partenza, di cui la CFA e i modelli di equazioni strutturali costituiscono uno sviluppo.

L'insegnamento pone una grande enfasi non solo sulla comprensione dei concetti teorici necessari per la costruzione e la validazione di uno strumento di misura in psicologia, ma anche sulla capacità di applicare tali concetti in situazioni concrete. Di conseguenza, la discussione dei concetti sarà sempre accompagnata da applicazioni pratiche. Tali applicazioni richiedono l'uso di un software. In questo insegnamento useremo R (R Core Team, 2021) quale linguaggio di programmazione probabilistica e, tra gli altri, il pacchetto `lavaan` che consente di svolgere le analisi statistiche della CFA e della EFA (Beaujean, 2014). La teoria classica dei test verrà descritta con riferimento al classico testo di Lord and Novick (1968). Questa dispensa, inoltre, segue da vicino la trattazione della CFA fornita nei testi di McDonald (2013) e di Brown (2015).

Trattando di argomenti avanzati, questo insegnamento presuppone la conoscenza di base dei concetti fondamentali della teoria delle probabilità; presuppone inoltre il possesso delle conoscenze di base necessarie per procedere all'utilizzo di R. Informazioni su tali argomenti sono forniti

nella dispensa di Psicometria (A.A. 2021-2022).

Corrado Caudek
Marzo 2022

Parte I

Il modello lineare

1

L'analisi di regressione

Conoscere l'analisi di regressione aiuta a capire la teoria classica dei test, l'analisi fattoriale e i modelli di equazioni strutturali. Sebbene le tecniche dell'analisi di regressione analizzino solo le variabili osservate, i principi della regressione costituiscono la base delle tecniche più avanzate che includono anche le variabili latenti.

1.1 Regressione bivariata

Il modello di regressione bivariata descrive l'associazione tra il valore atteso di $Y \mid x_i$ e x nei termini di una relazione lineare:

$$\mathbb{E}(Y \mid x_i) = \alpha + \beta x_i,$$

dove i valori x_i sono considerati fissi per disegno. Nel modello "classico", si assume che le distribuzioni $Y \mid x_i$ siano Normali con deviazione standard σ_ε.

Il significato dei coefficienti di regressione è semplice:

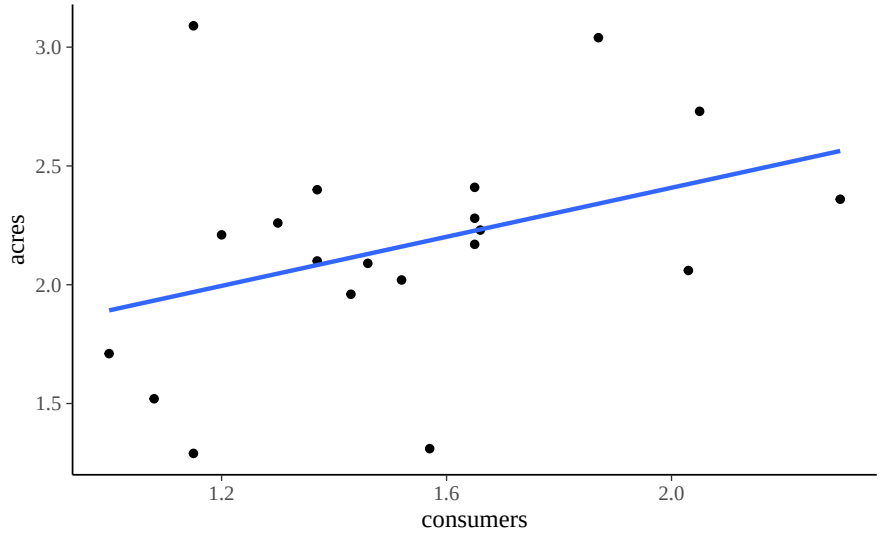
- α è il valore atteso di Y quando $X = 0$;
- β è l'incremento atteso nel valore atteso di Y quando X aumenta di un'unità.

Per fare un esempio, consideriamo i dati dell'antropologo Sahlins, il quale si è chiesto se esiste un'associazione tra l'ampiezza del clan (`consumers`) e l'area occupata da quel clan (`acres`) in una popolazione di cacciatori-raccoglitori. I dati sono i seguenti:

```
data(Sahlins)
head(Sahlins)
```

```
#>   consumers acres
#> 1      1.00  1.71
#> 2      1.08  1.52
#> 3      1.15  1.29
#> 4      1.15  3.09
#> 5      1.20  2.21
#> 6      1.30  2.26
```

```
Sahlins %>%
  ggplot(aes(x = consumers, y = acres)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)
```




```
fm <- lm(acres ~ consumers, data = Sahlins)
fm$coef
#> (Intercept)   consumers
#>      1.3756      0.5163
```

Dalla figura notiamo che, se `consumers` aumenta di un'unità (da 1.2 a 2.2), allora la retta di regressione (ovvero, il valore atteso di Y) aumenta di circa 0.5 punti – esattamente, aumenta di 0.5163 punti, come indicato

dalla stima del coefficiente β. L'interpretazione del coefficiente α è più problematica, perché non ha senso pensare ad un clan di ampiezza 0. Per affrontare questo problema, centriamo il predittore.

1.1.1 Regressori centrati

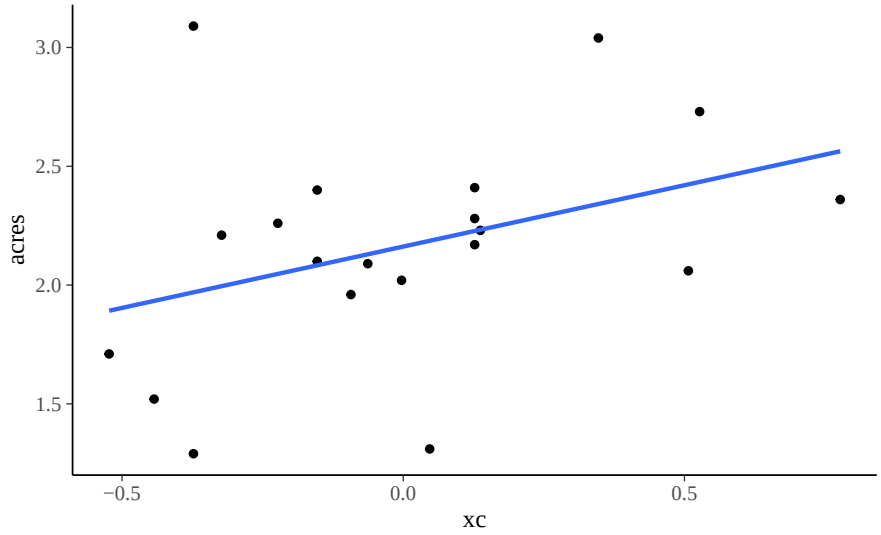
Esprimiamo la variabile `consumers` nei termini degli scarti dalla media:

```
Sahlins <- Sahlins %>%
  mutate(
    xc = consumers - mean(consumers)
  )
```

Svolgiamo nuovamente l'analisi di regressione con il nuovo predittore:

```
fm1 <- lm(acres ~ xc, data = Sahlins)
fm1$coef
#> (Intercept)          xc
#>      2.1620      0.5163
```

```
Sahlins %>%
  ggplot(aes(x = xc, y = acres)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)
```



La stima di β è rimasta invariata ma ora possiamo attribuire un significato alla stima di α: questo coefficiente indica il valore atteso della Y quando X assume il suo valore medio.

1.1.2 Minimi quadrati

La stima dei coefficienti del modello di regressione può essere effettuata in modi diversi: massima verosimiglianza o metodi bayesiani. Se ci limitiamo qui alla massima verosimiglianza possiamo semplificare il problema assumento che le distribuzioni condizionate $Y \mid x$ siano Normali. In tali circostanze, la stima dei coefficienti del modello di regressione può essere trovata con il metodo dei minimi quadrati.

In pratica, questo significa trovare i coefficienti a e b che minimizzano

$$SS_{\text{res}} = \sum (y_i - \hat{y}_i)^2,$$

con $\hat{y}_i = a + bx_i$.

Per fornire un'idea di come questo viene fatto, usiamo una simulazione. Per semplicità, supponiamo di conoscere $a = 1.3756445$ e di volere stimare b.

```
x <- Sahlins$consumers
y <- Sahlins$acres
```

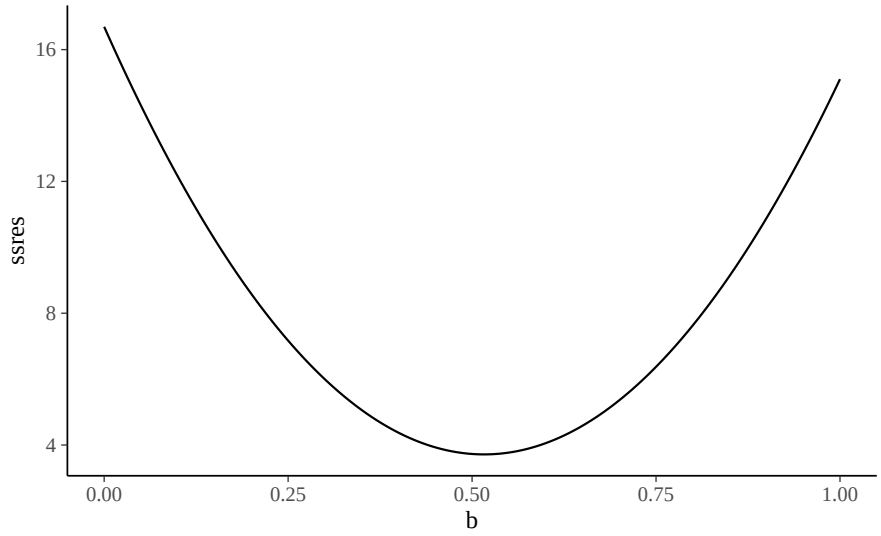
```
a <- 1.3756445

nrep <- 1e3
b <- seq(0, 1, length.out = nrep)

ssres <- rep(NA, nrep)
for (i in 1:nrep) {
  yhat <- a + b[i] * x
  ssres[i] <- sum((y - yhat)^2)
}
```

Un grafico di SS_{res} in funzione di b mostra che il valore b che minimizza SS_{res} corrisponde, appunto, a 0.5163.

```
tibble(b, ssres) %>%
  ggplot(aes(x = b, y = ssres)) +
  geom_line()
```



1.1.3 Relazione tra b e r

Un altro modo per interpretare b è quello di considerare la relazione tra la pendenza della retta di regressione e il coefficiente di correlazione:

$$b_X = r_{XY}\frac{S_X}{S_Y}$$

L'equazione precedente rende chiaro che, se i dati sono standardizzati, $b = r$.

Verifichiamo:

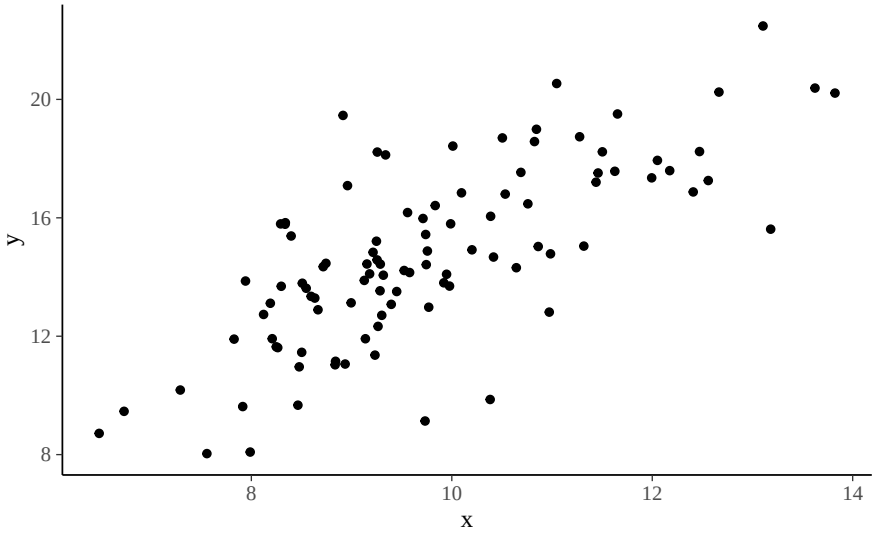
```
Sahlins %>%
  dplyr::select(acres, consumers) %>%
  cor()
#>            acres consumers
#> acres     1.0000    0.3757
#> consumers 0.3757    1.0000
```

```
fm2 <- lm(scale(acres) ~ scale(consumers), data = Sahlins)
fm2$coef
#>      (Intercept) scale(consumers)
#>        9.917e-17        3.757e-01
```

1.1.4 Attenuazione

Il fenomeno dell'attenuazione si verifica quando X viene misurato con una componente di errore. Esaminiamo la seguente simulazione.

```
set.seed(1234)
n <- 100
x <- rnorm(n, 10, 1.5)
y <- 1.5 * x + rnorm(n, 0, 2)
tibble(x, y) %>%
  ggplot(aes(x, y)) +
  geom_point()
```




```
sim_dat <- tibble(x, y)
fm <- lm(y ~ x, sim_dat)
fm$coef
#> (Intercept)           x
#>      0.4221      1.4652
```

Questi sono i coefficienti di regressione quando X è misurata senza errori.

```
sim_dat <- sim_dat %>%
  mutate(
    x1 = x + rnorm(n, 0, 2)
  )

fm1 <- lm(y ~ x1, sim_dat)
fm1$coef
#> (Intercept)          x1
#>      8.3872      0.6296
```

Aggiungendo una componente d'errore su X, la grandezza del coefficiente b diminuisce.

1.1.5 Coefficiente di determinazione

Tecnicamente, il coefficiente di determinazione è dato da:

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Al denominatore abbiamo la *devianza totale*, ovvero una misura della dispersione di y_i rispetto alla media $\bar{y}$. Al numeratore abbiamo una misura della dispersione del valore atteso della Y rispetto alla sua media. Il rapporto, dunque, ci dice qual è la quota della variabilità totale di Y che può essere predetta in base al modello lineare.

Per i dati di Sahlins abbiamo:

```
mod <- lm(acres ~ consumers, data = Sahlins)
a <- mod$coef[1]
b <- mod$coef[2]
yhat <- a + b * Sahlins$consumers
ss_tot <- sum((Sahlins$acres - mean(Sahlins$acres))^2)
ss_reg <- sum((yhat - mean(Sahlins$acres))^2)
r2 <- ss_reg / ss_tot
r2
#> [1] 0.1411
```

Verifichiamo:

```
summary(mod)
#>
#> Call:
#> lm(formula = acres ~ consumers, data = Sahlins)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.8763 -0.1873 -0.0211  0.2135  1.1206
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)    1.376      0.468    2.94   0.0088 **
#> consumers      0.516      0.300    1.72   0.1026
```

```
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.454 on 18 degrees of freedom
#> Multiple R-squared:  0.141,  Adjusted R-squared:  0.0934
#> F-statistic: 2.96 on 1 and 18 DF,  p-value: 0.103
```

Da cui deriva che R^2 è uguale al quadrato del coefficiente di correlazione:

```
cor(Sahlins$acres, Sahlins$consumers)^2
#> [1] 0.1411
```

1.1.6 Errore standard della regressione

L'errore standard della regressione è una stima della dispersione di $y \mid x_i$ nella popolazione. Non è altro che la deviazione standard dei residui

$$e = y_i - \hat{y}_i$$

che, al denominatore, riporta $n - 2$. La ragione è che, per calcolare $\hat{y}$, vengono "perduti" due gradi di libertà – il calcolo di $\hat{y}$ è basato sulla stima di due coefficienti: a e b.

```
e <- yhat - Sahlins$acres
(sum(e^2) / (length(Sahlins$acres) - 2)) %>%
  sqrt()
#> [1] 0.4543
```

Il valore trovato corrisponde a quello riportato nell'output di `lm()`.

1.2 Regressione multipla

Nella regressione multipla vengono utilizzati $k > 1$ predittori:

$$y_i = \alpha + \sum_{j=1}^{k} \beta_j x_i + \varepsilon_i.$$

L'interpretazione geometrica è simile a quella del modello bivariato. Nel caso di due predittori, il valore atteso della y può essere rappresentato da un piano; nel caso di $k > 2$ predittori, da un iper-piano. Nel caso di $k = 2$, tale piano è posto in uno spazio di dimensioni x_1, x_2 (che possiamo immaginare definire un piano orizzontale) e y (ortogonale a tale piano). La superficie piana che rappresenta $\mathbb{E}(y)$ è inclinata in maniera tale che l'angolo tra il piano e l'asse x_1 corrisponde a β_1 e l'angolo tra il piano e l'asse x_2 corrisponde a β_2.

1.2.1 Significato coefficienti parziali di regressione

Ai coefficienti parziali del modello di regressione multipla possiamo assegnare la seguente interpretazione:

Il coefficiente parziale di regressione β_j rappresenta l'incremento atteso della y se x_j viene incrementata di un'unità, tenendo costante il valore delle altre variabili indipendenti.

Un modo per interpretare la locuzione "al netto dell'effetto delle altre variabili indipendenti" è quello di esaminare la relazione tra la y parzializzata e la x_j parzializzata. In questo contesto, parzializzare significa decomporre una variabile di due componenti: una componente che è linearmente predicibile da una o più altre variabili e una componente che è linearmente incorrelata con tali varibili "terze".

Eseguiamo questa "depurazione" dell'effetto delle variabili "terze" sia sulla y sia su x_j. A questo punto possiamo esaminare la relazione bivariata che intercorre tra la componente della y linearmente indipendente dalle variabili "terze" e la componente della x_j linearmente indipendente dalle variabili "terze". Il coefficiente di regressione bivariato così ottenuto è identico al coefficiente parziale di regressione nel modello di regressione multipla. Possiamo così ottenere un'interpretazione di β_j.

Esaminiamo un caso concreto.

```
d <- rio::import(
  here::here("data", "kidiq.dta")
)
```

```
glimpse(d)
#> Rows: 434
#> Columns: 5
#> $ kid_score <dbl> 65, 98, 85, 83, 115, 98, 69, 106, 1~
#> $ mom_hs    <dbl> 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,~
#> $ mom_iq    <dbl> 121.12, 89.36, 115.44, 99.45, 92.75~
#> $ mom_work  <dbl> 4, 4, 4, 3, 4, 1, 4, 3, 1, 1, 1, 4,~
#> $ mom_age   <dbl> 27, 25, 27, 25, 27, 18, 20, 23, 24,~
```

```
fm <- lm(kid_score ~ mom_iq + mom_work + mom_age + mom_hs, data = d)
fm$coef
#> (Intercept)      mom_iq    mom_work     mom_age
#>     20.8226      0.5621      0.1337      0.2199
#>      mom_hs
#>      5.5612
```

Eseguiamo la parzializzazione di y in funzione delle variabili `mom_work`, `mom_age` e `mom_hs`:

```
fm_y <- lm(kid_score ~ mom_work + mom_age + mom_hs, data = d)
```

Lo stesso per `mom_iq`:

```
fm_x <- lm(mom_iq ~ mom_work + mom_age + mom_hs, data = d)
```

Esaminiamo ora la regressione bivariata tra le componenti parzializzate della y e di x_j:

```
mod <- lm(fm_y$residuals ~ fm_x$residuals)
mod$coef
#>    (Intercept) fm_x$residuals
#>     -1.652e-15      5.621e-01
```

Si vede come il coefficiente di regressione bivariato risulta identico al corrispondente coefficiente parziale di regressione.

1.2.2 Relazioni causali

Un altro modo per interpretare i coefficienti parziali di regressione è nell'ambito dei quelli che vengono chiamati i *path diagrams*. I diagrammi di percorso, che tratteremo in seguito e qui solo anticipiamo, descrivono le relazioni "causali" tra variabili: le variabili a monte del diagramma di percorso indicono le "cause" esogene e le variabili a valle indicano gli effetti, ovvero le variabili endogene. I coefficienti di percorso rappresentati graficamente come frecce orientate corrispondono all'effetto *diretto* sulla variabile verso cui punta la freccia della variabile a monte della freccia. Tali coefficienti di percorso non sono altro che i coefficienti parziali di regressione del modello di regressione multipla. In questo contesto, indicano l'effetto atteso *diretto* sulla variabile endogena dell'incremento di un'unità della variabile esogena, lasciano immutate tutte le altre relazioni strutturali del modello.

Usiamo la funzione `sem()` del pacchetto `lavaan` per definire il modello rappresentato nel successivo diagramma di percorso:

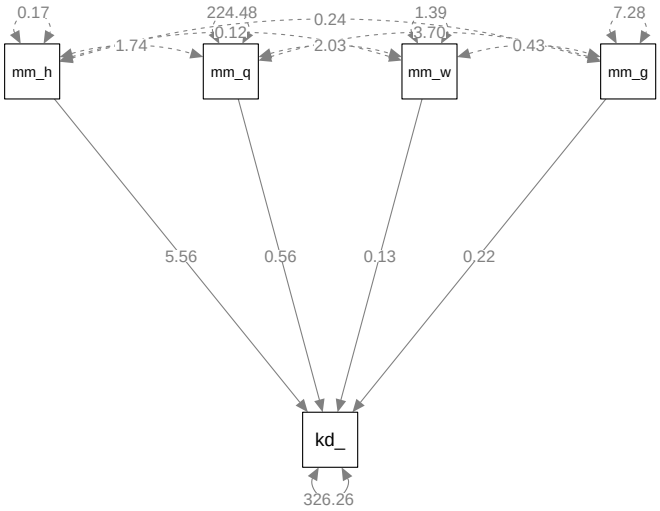
```
model <- "
  kid_score ~ mom_hs + mom_iq + mom_work + mom_age
"
```

Adattiamo il modello ai dati

```
fit <- sem(model, data = d)
```

Il diagramma di percorso si ottiene con le seguenti istruzioni:

```
semPaths(
  fit, "est",
  posCol = c("black"),
  edge.label.cex = 0.9,
  sizeMan = 7,
  what = "path"
)
```



Come indicato nel diagramma, l'effetto diretto di `mom_iq` su `kid_score` è identico al corrispondente coefficiente parziale di regressione.

1.2.3 Errore di specificazione

Spiritosamente chiamato "heartbreak of L.O.V.E." (Left-Out Variable Error; Mauro (1990)), l'errore di specificazione è una caratteristica fondamentale dei modelli di regressione che deve sempre essere tenuta a mente quando interpretiamo i risultati di questa analisi statistica. L'errore di specificazione si verifica quando escludiamo dal modello di regressione una variabile che

- è associata con altre variabili nel modello,
- ha un effetto diretto sulla y.

Come conseguenza dell'errore di specificazione, la direzione e il segno dei coefficienti parziali di regressione risultano sistematicamente distorti.

Consideriamo un esempio con dati simulati nei quali immaginiamo che la prestazione sia positivamente associata alla motivazione e negativamente associata all'ansia. Immaginiamo inoltre che vi sia una correlazione positiva tra ansia a motivazione. Ci chiediamo cosa succede al coefficiente parziale della variabile "motivazione" se la variabile "ansia" viene esclusa dal modello di regressione.

```
set.seed(123)
anxiety <- rnorm(n, 10, 1.5)
motivation <- 4.0 * anxiety + rnorm(n, 0, 3.5)
cor(anxiety, motivation)
#> [1] 0.8435
```

```
performance <- 0.5 * motivation - 5.0 * anxiety + rnorm(n, 0, 3)
```

```
sim_dat2 <- tibble(performance, motivation, anxiety)
fm1 <- lm(performance ~ motivation + anxiety, sim_dat2)
coef(fm1)
#> (Intercept)  motivation     anxiety
#>      3.0686      0.5204     -5.3480
```

```
fm2 <- lm(performance ~ motivation, sim_dat2)
summary(fm2)
#>
#> Call:
#> lm(formula = performance ~ motivation, data = sim_dat2)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -12.110  -3.509   0.334   2.940  16.725
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -11.7178     3.1592   -3.71  0.00034 ***
#> motivation   -0.4610     0.0777   -5.93  4.5e-08 ***
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4.87 on 98 degrees of freedom
#> Multiple R-squared:  0.264,  Adjusted R-squared:  0.257
#> F-statistic: 35.2 on 1 and 98 DF,  p-value: 4.51e-08
```

Il risultato prodotto dal modello di regressione è sbagliato: come conseguenza dell'errore di specificazione, il segno del coefficiente parziale di regressione della variabile "motivazione" è negativo, anche se nel vero modello di regressione tale coefficiente ha il segno opposto. Quindi, se noi interpretassimo il coefficiente parziale ottenuto in termini casuali, saremmo portati a concludere che la motivazione fa diminuire la prestazione anche se, in realtà (nel modello generatore dei dati), è vero l'opposto.

1.2.4 Soppressione

Le conseguenze dell'errore di specificazione sono chiamate "soppressione" (*suppression*). In generale, si ha soppressione quando (1) il valore assoluto del peso beta di un predittore è maggiore di quello della sua correlazione bivariata con il criterio o (2) i due hanno segni opposti.

- L'esempio descritto sopra è un caso di *soppressione negativa*, dove il predittore ha correlazioni bivariate positive con il criterio, ma si riceve un peso beta negativo nell'analisi di regressione multipla.
- Un secondo tipo di soppressione è la *soppressione classica*, in cui un predittore non è correlato al criterio ma riceve un peso beta diverso da zero se un altro predittore viene controllato.
- C'è anche la *soppressione reciproca* che può verificarsi quando due variabili sono correlate positivamente con il criterio ma negativamente tra loro.

1.2.5 Stepwise regression

Un'implicazione della soppressione è che i predittori non dovrebbero essere selezionati in base ai valori delle correlazioni bivariate con il criterio. Queste associazioni di ordine zero non controllano gli effetti degli altri predittori, quindi i loro valori possono essere fuorvianti rispetto ai coefficienti di regressione parziale per le stesse variabili. Per lo stesso motivo, il fatto che le correlazioni bivariate con il criterio siano statisticamente significative o meno è irrilevante per quanto riguarda la selezione dei predittori. Sebbene le procedure informatiche di regressione rendano facile tali processi di selezione dei predittori, i ricercatori dovrebbero evitare di usare tali metodi. Il rischio è che anche piccole, ma non rilevate, nonlinearità o effetti indiretti tra i predittori possano seriamente distocere i coefficienti di regressione parziale. È meglio selezionare giudiziosamente il minor numero di predittori sulla base di ragioni teoriche o dei risultati di ricerche precedenti.

Una volta selezionati, ci sono due modi di base per inserire i predittori nell'equazione di regressione: uno consiste nell'inserire tutti i predittori contemporaneamente. L'altro è inserirli nel corso di una serie di passaggi, ovvero mediante usando una procedura sequenziale. L'ordine di ingresso può essere determinato in base a uno di due diversi standard: teorici (razionali) o empirici (statistici). Lo standard razionale corrisponde alla regressione gerarchica, in cui si comunica al computer un ordine fisso per inserire i predittori. Ad esempio, a volte le variabili demografiche vengono inserite nel primo passaggio, quindi nel secondo passaggio viene inserita una variabile psicologica di interesse. Questo ordine non solo controlla le variabili demografiche ma permette anche di valutare il potere predittivo della variabile psicologica, al di là di quello delle semplici variabili demografiche. Quest'ultimo può essere stimato come l'aumento della correlazione multipla al quadrato, o ΔR^2, da quella della fase 1 con solo predittori demografici a quella della fase 2 con tutti i predittori nell'equazione di regressione.

Un esempio di standard statistico è la regressione *stepwise*, in cui il computer seleziona l'inserimento dei predittori in base esclusivamente alla significatività statistica; cioè, viene chiesto: quale predittore, se inserito nell'equazione, avrebbe il valore_p più piccolo per il test del suo coefficiente di regressione parziale? Dopo la selezione, i predittori in una fase successiva possono essere rimossi dall'equazione di regressione in base ai loro valori-p (ad esempio, se $p \geq .05$). Il processo stepwise si interrompe quando, aggiungendo più predittori, ΔR^2 non migliora. Varianti della regressione stepwise includono *forward inclusion*, in cui i predittori selezionati non vengono successivamente rimossi dal modello, e *backward elimination*, che inizia con tutti i predittori nel modello per poi rimuoverne alcuni in passi successivi. I problemi relativi ai metodi stepwise sono così gravi da essere effettivamente banditi in alcuni giornali. Un problema è che fanno leva su risultati che si ottengono per caso, in dipendenza delle idiosincrasie del campione (quindi, non replicabili).

In secondo luogo, una volta che un insieme finale di predittori selezionati razionalmente è stato inserito nell'equazione di regressione, tali predittori non dovrebbero essere successivamente rimossi se i loro coefficienti di regressione non sono statisticamente significativi: il ricercatore non dovrebbe sentirsi in dovere di lasciar perdere ogni predittore che non risulta statisticamente significativo. In campioni piccoli, la potenza dei test di significatività è bassa e la rimozione di un predittore non significativo può alterare sostanzialmente la soluzione. Se c'è una buona ragione

per includere un predittore, allora è meglio lasciarlo nel modello, fino a prova contraria.

Parte II

La teoria classica dei test

2

Fondamenti teorici

2.1 Valutazione psicometrica come ragionamento inferenziale

In apparenza, i test psicometrici sono solo dei test. Somministriamo un test, otteniamo un punteggio ed è naturale pensare che sia tutto lì. Nonostante le apparenze, la valutazione psicologica e neuropsicologica non consiste soltanto nell'assegnare di punteggi: si tratta di ragionare su ciò che osserviamo di quello che le persone dicono, fanno o producono, in maniera tale da giungere a delle concezioni più ampie di tali persone a proposito di aspetti che non abbiamo – e spesso non possiamo – osservare. Più specificamente, possiamo considerare la valutazione psicologica e neuropsicologica come un esempio di ragionamento che fa uso di modelli probabilistici per giungere a delle spiegazioni, previsioni o conclusioni.

I dati osservati diventano un'evidenza quando sono ritenuti rilevanti per l'inferenza desiderata attraverso l'instaurazione di relazioni tra i dati e l'obiettivo dell'inferenza. Spesso utilizziamo dati provenienti da più fonti. Queste possono essere di tipo simile (ad esempio, item di test aventi lo stesso formato) o di tipo molto diverso (ad esempio, il curriculum di un richiedente oltre al colloquio, la storia medica della famiglia di un paziente, ...). Le evidenze possono essere contraddittorie (ad esempio, uno studente riesce a svolgere un compito difficile ma fallisce in un uno facile) e quasi sempre non sono del tutto conclusive.

Queste caratteristiche hanno due implicazioni. In primo luogo, è difficile capire cosa le evidenze implicano. I processi inferenziali sono sempre complessi. In secondo luogo, a causa della natura non conclusiva delle evidenze disponibili, non siamo mai del tutto certi delle nostre inferenze. Per affrontare tale incertezza, la teoria psicometria ci fornisce gli strumenti che ci possono aiutare nel processo inferenziale, dai dati disponibili alle decisioni che prendiamo.

Un secolo fa, la relazione tra prestazioni osservate, da un lato, e l'abilità inosservabile del rispondente, dall'altro, iniziò a essere formalizzata nei termini dell'*errore di misurazione*. Gulliksen (1961) ha descritto "il problema centrale della teoria dei test" come "la relazione tra l'abilità dell'individuo e il suo punteggio osservato sul test" (p. 101). Tale caratterizzazione è valida ancora oggi, con una definizione opportunamente ampia di "abilità" e di "punteggio sul test" che sia in grado di comprendere le diverse forme di assessment psicologico e neuropsicologico. Comprendere e essere in grado di rappresentare la relazione tra le prestazioni osservate e la capacità soggiacente è dunque fondamentale per le forme di ragionamento che vengono impiegate nella valutazione psicologica e neuropsicologica.

Come risultato dell'errore di misurazione, i ragionamenti che compiamo nella valutazione psicologica e neuropsicologica costituiscono un esempio di ragionamento in condizioni di incertezza. A causa della natura imperfetta della misurazione e dell'incompletezza dell'informazione disponibile, le nostre inferenze sono incerte e possono essere sempre invalidate o riviste. Ragionare da ciò che è parziale (ciò che vediamo uno paziente dire, fare o produrre) a ciò che è generale (la "vera" abilità del paziente) è necessariamente incerto, e le nostre inferenze o conclusioni sono sempre prone ad errori.

Quali strumenti devono essere impiegati per affrontare la nostra incertezza sulla relazione che intercorre tra prestazioni osservate e abilità soggiacenti? Secondo Lewis, molti dei progressi nella teoria psicometrica sono resi possibili "trattando lo studio della relazione tra le risposte agli item di un test e il tratto ipotizzato di un individuo come problema di inferenza statistica" (Lewis, 1986). Una connessione diretta tra errore di misura e approccio probabilistico è stata anche proposta da Samejima: "There may be an enormous number of factors eliciting a student's specific overt reactions to a stimulus, and, therefore, it is suitable, even necessary, to handle the situation in terms of the probabilistic relationship between the two" (Samejima, 1983).

Questo punto di vista è diventato quello dominante nella psicometria moderna e sottolinea l'utilità di utilizzare il linguaggio e gli strumenti della teoria della probabilità per comunicare il carattere parziale dei dati di cui dispone lo psicologo e l'incertezza delle inferenze che ne derivano.

I reattivi psicologici possono essere costruiti e la validati mediante vari approcci probabilistici: la Teoria Classica dei test (*classical test theory*, in

breve CTT) e la teoria di risposta all'item (*item response theory*, in breve IRT) sono quelli più noti. Recentemente, il problema della valutazione psicologica è stato anche formulato in un'ottica bayesiana. In questo insegnamento esamineremo gli approcci della CTT e dell'IRT, ma non quello bayesiano.

2.2 La Teoria Classica

La CTT nasce alla fine dell'Ottocento (Alfred Binet e altri, 1894) allo scopo di studiare l'attendibilità e la validità dei risultati dei questionari utilizzati per valutare le caratteristiche psico-sociali, non direttamente osservabili, delle persone esaminate. L'impiego su vasta scala e lo sviluppo della CTT ha inizio negli anni Trenta, anche se il modello formale su cui tale teoria si basa viene proposta da Spearman all'inizio del Novecento (Spearman, 1904). La tecnica dell'analisi fattoriale esplorativa (*Exploratory Factor Analysis*, EFA), verrà poi affinata da Thurstone (1947) alla fine della seconda guerra mondiale. Tra la fine degli anni '60 e gli inizi degli anni '70, Jöreskog (1969) sviluppa l'analisi fattoriale confermativa (*Confirmatory Factor Analysis*, CFA). Negli anni '70, l'analisi fattoriale viene integrata con la path analysis nel lavoro di Jöreskog (1978) che dà origine ai modelli di equazioni strutturali (*Structural Equation Modeling*, SEM).

Iniziamo qui ad esaminare queste tecniche psicometriche prendendo in esame, per prima, la teoria classica dei test. Seguiremo la trattazione proposta da Lord and Novick (1968).

L'equazione fondamentale alla quale si riconduce la teoria classica dei test è quella che ipotizza una relazione lineare e additiva tra il punteggio osservato di un test (X), la misura della variabile latente (T) e la componente casuale dell'errore (E). Un punto cruciale nella CTT è l'entità della varianza dell'errore. Minore è la varianza dell'errore, più accuratamente il punteggio reale viene riflesso dai nostri punteggi osservati. In un mondo perfetto, tutti i valori di errore sarebbero uguali a 0. Cioè, ogni partecipante otterrebbe il punteggio esatto. Questo però non è possibile. Pertanto, abbiamo una certa varianza negli errori. La corrispondente deviazione standard di tali errori ha il un nome: si chiama *errore standard di misurazione*, indicato da σ_E. Uno dei principali obiettivi della CTT è

quello di ottenere una stima di σ_E.

2.3 Le due componenti del punteggio osservato

CTT si occupa delle relazioni tra X, T ed E. La CTT si basa su un modello relativamente semplice in cui il punteggio osservato, il punteggio vero (cioè l'abilità inosservabile del rispondente) e l'errore aleatoria di misurazione sono legati da una relazione lineare. Indicati con $T_{\nu j}$ (*true score*) l'abilità latente da misurare dell'individuo ν nella prova j, con $X_{\nu j}$ la variabile osservata (*observed score*) per l'individuo ν nella prova j e con $E_{\nu j}$ l'errore aleatorio di misurazione, il modello si rappresenta con

$$X_{\nu j} = T_{\nu} + E_{\nu j}.$$

Dunque, in base alla (2.3) il punteggio osservato $X_{\nu j}$ differisce da quello vero $T_{\nu j}$ a causa di una componente di errore aleatoria $E_{\nu j}$. Uno degli obiettivi centrali della CTT è quello di quantificare l'entità di tale errore. Vedremo come questa quantificazione verrà fornita nei termini dell'attendibilità del test. L'attendibilità (o affidabilità) rappresenta l'accuratezza con cui un test può misurare il punteggio vero (Coaley, 2014):

- Se l'attendibilità è grande, σ_E è piccolo: X ha un piccolo errore di misurazione e sarà vicino a T.
- Se l'attendibilità è piccola, σ_E è grande: X presenta un grande errore di misurazione e si discosterà molto da T.

2.3.1 Il punteggio vero

La @ref(eq:observed-true_plus-error) ci dice che il punteggio osservato è dato dalla somma di due componenti: una componente sistematica (il punteggio vero) e una componente aleatoria (l'errore di misurazione). Ma che cos'è il punteggio vero? La CTT considera un reattivo psicologico come una selezione aleatoria di item da un universo/popolazione di item attinenti al costrutto da misurare (Nunnally, 1994; Kline, 2013). Se il reattivo psicologico viene concepito in questo modo, il punteggio vero diventa il punteggio che un rispondente otterrebbe se fosse misurato su tutto l'universo degli item proprio del costrutto in esame. L'errore di

misurazione riflette dunque il grado in cui gli item che costituiscono il test non riescono a rappresentare l'intero universo degli item attinenti al costrutto.

In maniera equivalente, il punteggio vero può essere concepito come il punteggio non "distorto" da componenti estranee al costrutto, ovvero da effetti di apprendimento, fatica, memoria, motivazione, eccetera. Essendo concepita come del tutto aleatoria (ovvero, priva di qualunque natura sistematica), la componente aleatoria non introduce dei bias nella tendenza centrale della misurazione.

Il punteggio vero è concepito come un punteggio inosservabile che corrisponde al valore atteso di infinite realizzazioni del punteggio ottenuto:

$$T = \mathbb{E}(X) \equiv \mu_X \equiv \mu_T.$$

In altri termini, secondo la definizione di Lord e Novick (1968), e facendo riferimento a alla seconda definizione presentata sopra, il punteggio vero è concepito come la media dei punteggi che un soggetto otterrebbe se il test venisse somministrato ripetutamente nelle stesse condizioni, in assenza di effetti di apprendimento e/o fatica.

2.3.2 Somministrazioni ripetute

Nella formulazione del modello della CTT si possono distinguere due tipi di esperimenti aleatori: uno che considera l'unità di osservazione (l'individuo) come campionaria, l'altro che considera il punteggio, per un determinato individuo, come campionario. Un importante risultato è dato dall'unione dei due esperimenti, ovvero dalla dimostrazione che i risultati della CTT, la quale è stata sviluppata ipotizzando ipotetiche somministrazioni ripetute del test allo stesso individuo sotto le medesime condizioni, si generalizzano al caso di una singola somministrazione del test ad un campione di individui (Allen and Yen, 2001). In base a questo risultato, se consideriamo la somministrazione del test ad una popolazione di individui, allora diventa più facile dare un contenuto empirico alle quantità della CTT:

- σ^2_X è la varianza del punteggio osservato nella popolazione,
- σ^2_T è la varianza dei punteggio vero nella popolazione,
- σ^2_E è la varianza della componente d'errore nella popolazione.

2.3.3 Le assunzioni sul punteggio ottenuto

La CTT *assume* che la media del punteggio osservato X sia uguale alla media del punteggio vero,

$$\mu_X \equiv \mu_T,$$

in altri termini, assume che il punteggio osservato fornisca una stima statisticamente corretta dell'abilità latente (punteggio vero). In pratica, il punteggio osservato non sarà mai uguale all'abilità latente, ma corrisponde solo ad uno dei possibili punteggi che il soggetto può ottenere, subordinatamente alla sua abilità latente. L'errore della misura è la differenza tra il punteggio osservato e il punteggio vero: $E \equiv X - T$.

In base all'assunzione secondo cui il valore atteso dei punteggi è uguale alla media del valore vero, segue che

$$\mathbb{E}(E) = \mathbb{E}(X - T) = \mathbb{E}(X) - \mathbb{E}(T) = \mu_T - \mu_T = 0,$$

ovvero, il valore atteso degli errori è uguale a zero.

2.4 L'errore standard della misurazione σ_E

La radice quadrata di σ_E^2, ovvero la deviazione standard degli errori, è la quantità fondamentale della CTT ed è chiamata *errore standard della misurazione*. La stima dell'errore standard della misurazione costituisce uno degli obiettivi più importanti della CTT. Ricordiamo che la deviazione standard è simile (non identica) alla media del valore assoluto degli scarti dei valori di una distribuzione dalla media. Possiamo dunque utilizzare questa proprietà per descrivere il modo in cui la CTT interpreta σ_E.

L'*errore standard della misurazione* σ_E ci dice qual è, approssimativamente, la variazione attesa del punteggio osservato, se il test venisse somministrato un'altra volta al rispondente nelle stesse condizioni.

2.5 Assiomi della Teoria Classica

La CTT *assume* che gli errori siano delle variabili aleatorie incorrelate tra loro

$$\rho(E_i, E_k \mid T) = 0, \qquad \text{con } i \neq k,$$

e incorrelate con il punteggio vero,

$$\rho(E, T) = 0,$$

le quali seguono una distribuzione gaussiana con media zero e deviazione standard pari a σ_E:

$$E \sim \mathcal{N}(0, \sigma_E).$$

La quantità σ_E è detta errore standard della misurazione.

Sulla base di tali assunzioni la CTT deriva la formula dell'attendibilità di un test. Si noti che le assunzioni della CTT hanno una corrispondenza puntuale con le assunzioni su cui si basa il modello di regressione lineare.

2.6 L'attendibilità del test

Il concetto di attendibilità è strettamente legato alla riproducibilità della misurazione: si riferisce al grado di stabilità, di coerenza interna e di precisione di una procedura di misurazione. Affinché una misurazione psicologica sia utile, deve produrre lo stesso risultato se viene applicata ripetutamente un determinato rispondente. Altri termini che vengono usati sono: affidabilità, costanza e credibilità.

Vedremo nel seguito come il coefficiente di attendibilità fornisce una stima della quota della varianza del punteggio osservato che può essere attribuita all'abilità latente ("punteggio vero", cioè privo di errore di misurazione). In generale, un coefficiente di attendibilità maggiore di 0.80 viene ritenuto soddisfacente perché indica che l'80% o più della varianza

dei punteggi ottenuti è causata da ciò che il test intende misurare, anziché dall'errore di misurazione.

Per definire l'attendibilità, la CTT si serve di due quantità:

- la varianza del punteggio osservato,
- la correlazione tra punteggio osservato e punteggio vero.

Vediamo come queste quantità possano essere ottenute sulla base delle assunzioni del modello statistico che sta alla base della CTT.

2.6.1 La varianza del punteggio osservato

La varianza del punteggio osservato X è uguale alla somma della varianza del punteggio vero e della varianza dell'errore di misurazione.

Dimostrazione. La varianza del punteggio osservato è uguale a

$$\sigma_X^2 = \mathbb{V}(T + E) = \sigma_T^2 + \sigma_E^2 + 2\sigma_{TE}.$$

Dato che $\sigma_{TE} = \rho_{TE}\sigma_T\sigma_E = 0$, in quanto $\rho_{TE} = 0$, ne segue che

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2.$$

□

2.6.2 La covarianza tra punteggio osservato e punteggio vero

La covarianza tra punteggio osservato X e punteggio vero T è uguale alla varianza del punteggio vero.

Dimostrazione.

$$\begin{aligned}
\sigma_{XT} &= \mathbb{E}(XT) - \mathbb{E}(X)\mathbb{E}(T) \\
&= \mathbb{E}[(T + E)T] - \mathbb{E}(T + E)\mathbb{E}(T) \\
&= \mathbb{E}(T^2) + \underbrace{\mathbb{E}(ET)}_{=0} - [\mathbb{E}(T)]^2 - \underbrace{\mathbb{E}(E)}_{=0}\,\mathbb{E}(T) \\
&= \mathbb{E}(T^2) - [\mathbb{E}(T)]^2 \\
&= \sigma_T^2.
\end{aligned}$$

□

Da ciò segue che la correlazione tra punteggio osservato X e punteggio vero T è uguale al rapporto tra la deviazione standard del punteggio vero e la deviazione standard del punteggio osservato:

$$\rho_{XT} = \frac{\sigma_{XT}}{\sigma_X \sigma_T} = \frac{\sigma_T^2}{\sigma_X \sigma_T} = \frac{\sigma_T}{\sigma_X}. \tag{2.1}$$

2.6.3 Definizione e significato dell'attendibilità

La CTT definisce attendibilità di un test (o di un item) come il quadrato della correlazione tra punteggio osservato X e punteggio vero T, ovvero come il rapporto tra la varianza del punteggio vero e la varianza del punteggio osservato:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}.$$

Questa è la quantità fondamentale della CTT e misura il grado di variazione del punteggio vero rispetto alla variazione del punteggio osservato[^2].

Dato che $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$, in base alla (2.6.3) possiamo scrivere

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}. \tag{2.2}$$

Questo significa che il coefficiente di attendibilità assume valore 1 se la varianza degli errori σ_E^2 è nulla e assume valore 0 se la varianza degli errori è uguale alla varianza del punteggio osservato. Ciò significa che il coefficiente di attendibilità è un numero contenuto nell'intervallo compreso tra 0 e 1.

2.7 Attendibilità e modello di regressione lineare

Il modello di regressione lineare sta alla base della CTT. Infatti si può dire che tutte le proprietà della CTT che abbiamo discusso in precedenza non sono altro che le caratteristiche di un modello di regressione lineare nel quale

- la variabile dipendente è costituita dai punteggi osservati X, e
- la variabile indipendente corrisponde ai punteggi veri T.

Se rappresentiamo la CTT in questo modo, il coefficiente di attendibilità $\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X}$ non diventa altro che la quota di varianza del punteggio osservato X che viene spiegata dal punteggio vero T in base ad un modello lineare con pendenza unitaria e intercetta nulla. Nei termini di una tale rappresentazione, il coefficiente di attendibilità misura la forza della relazione lineare tra X e T e corrisponde al coefficiente di determinazione del seguente modello di regressione:

$$X = 0 + 1 \cdot T + E.$$

2.7.1 Simulazione

Per dare un contenuto concreto alle affermazioni precedenti, consideriamo la seguente simulazione svolta in . In tale simulazione il punteggio vero T e l'errore E verranno creati in modo tale da soddisfare i due vincoli della CTT: T e E saranno delle variabili gaussiane e tra loro incorrelate. Nella simulazione generiamo 100 coppie di valori X e T con i seguenti parametri: $T \sim \mathcal{N}(\mu_T = 12, \sigma^2_T = 6)$, $E \sim \mathcal{N}(\mu_E = 0, \sigma^2_T = 3)$. A tale fine usiamo le seguenti istruzioni:

```
set.seed(123)
library("MASS")
n <- 100
Sigma <- matrix(c(6, 0, 0, 3), byrow = TRUE, ncol = 2)
Sigma
#>      [,1] [,2]
#> [1,]    6    0
#> [2,]    0    3
```

```
mu <- c(12, 0)
mu
#> [1] 12  0
```

```
Y <- mvrnorm(n, mu, Sigma, empirical = TRUE)
T <- Y[, 1]
E <- Y[, 2]
```

Le istruzioni precedenti creano un insieme di valori tali per cui le medie e la matrice di varianze-covarianze assumono esattamente i valori indicati. Possiamo dunque immaginare tale insieme di dati come la nostra "popolazione".

Secondo la CTT, il punteggio osservato è $X = T + E$. Simuliamo dunque il punteggio osservato X nel modo seguente:

```
X <- T + E
```

Le prime 6 osservazioni così ottenute sono:

```
head(cbind(T, E, X))
#>           T       E      X
#> [1,] 11.148 -1.5708  9.577
#> [2,] 13.138 -0.3335 12.804
#> [3,] 10.391  2.5457 12.937
#> [4,] 11.452 -0.1955 11.257
#> [5,]  9.978 -0.4920  9.486
#> [6,] 10.730  2.9609 13.691
```

Un diagramma di dispersione è fornito nella figura seguente:

```
tibble(X, T) %>%
  ggplot(aes(X, T)) +
  geom_point()
```

Secondo la CTT, il valore atteso di T è uguale al valore atteso di X. Verifichiamo questa assunzione della CTT nei nostri dati:

```
mean(T)
#> [1] 12
mean(X)
#> [1] 12
```

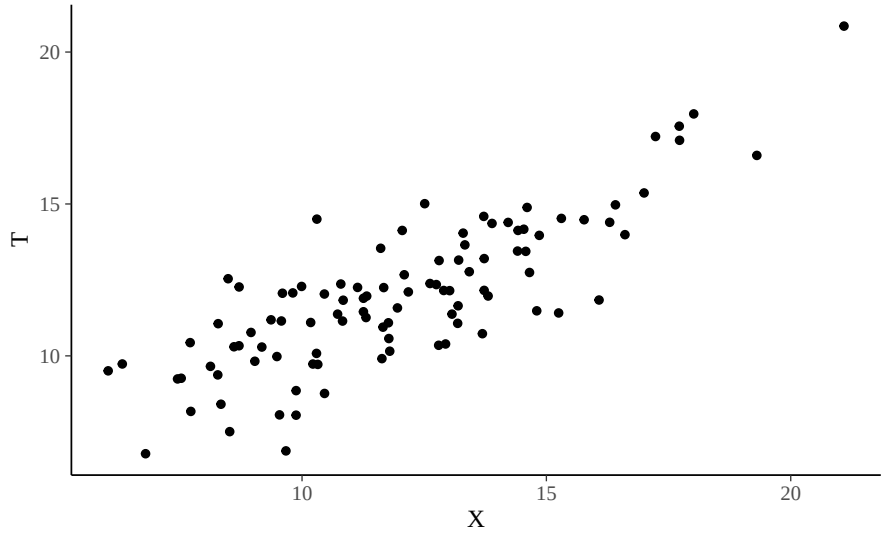


Figura 2.1: *Simulazione della relazione tra punteggio osservato e punteggio vero per 100 individui in base alle assunzioni della CTT.*

L'errore deve avere media zero, varianza σ_E^2 e deve essere incorrelato con T:

```
mean(E)
#> [1] 4.061e-18
var(E)
#> [1] 3
cor(T, E)
#> [1] -1.947e-16
```

Ricordiamo che la radice quadrata della varianza degli errori è chiamata errore standard della misurazione, σ_E. La quantità $\sqrt{\sigma_E^2}$ fornisce una misura della dispersione del punteggio osservato attorno al valore vero, nella condizione ipotetica di effettuare ripetute somministrazioni del test.

```
sqrt(3)
#> [1] 1.732
```

Dato che T e E sono incorrelati, ne segue che la varianza del punteggio osservato X è uguale alla somma della varianza del punteggio vero T e

della varianza degli errori E:

```
var(X)
#> [1] 9
var(T) + var(E)
#> [1] 9
```

La varianza del punteggio vero T è uguale alla covarianza tra il punteggio vero T e il punteggio osservato X:

```
cov(T, X)
#> [1] 6
var(T)
#> [1] 6
```

La correlazione tra il punteggio osservato e il punteggio vero è uguale al rapporto tra la deviazione standard del punteggio vero e la deviazione standard del punteggio osservato:

```
cor(X, T)
#> [1] 0.8165
sd(T) / sd(X)
#> [1] 0.8165
```

Focalizziamoci ora sull'attendibiiltà. Per la CTT, l'attendibilità è uguale al quadrato del coefficiente di correlazione tra il punteggio vero T e il punteggio osservato X:

```
cor(X, T)^2
#> [1] 0.6667
```

La motivazione di questa simulazione è quella di mettere in relazione il coefficiente di attendibilità, calcolato con le formule della CTT, con il modello di regressione lineare. Analizziamo dunque i dati della simulazione mediante il seguente modello di regressione lineare:

$$X = a + bT + E$$

```
fm <- lm(X ~ T)
summary(fm)
#>
#> Call:
#> lm(formula = X ~ T)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -4.197 -1.101  0.052  1.155  4.239
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 8.53e-15   8.75e-01       0        1
#> T           1.00e+00   7.14e-02      14   <2e-16 ***
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.74 on 98 degrees of freedom
#> Multiple R-squared:  0.667,  Adjusted R-squared:  0.663
#> F-statistic:  196 on 1 and 98 DF,  p-value: <2e-16
```

Si noti che la retta di regressione ha intercetta 0 e pendenza 1. Questo è coerente con l'assunzione $\mathbb{E}(X) = \mathbb{E}(T)$. Ma il risultato più importante di questa simulazione è il seguente: il coefficiente di determinazione ($R^2 = 0.67$) del modello di regressione $X = 0 + 1 \times T + E$ è identico al coefficiente di attendibilità che abbiamo calcolato con la formula $\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X}$:

```
var(T) / var(X)
#> [1] 0.6667
```

Ciò ci consente di attribuire al coefficiente di attendibilità la seguente interpretazione: l'attendibilità di un test non è altro che la quota di varianza del punteggio osservato X che viene spiegata dalla regressione di X sul punteggio vero T in un modello di regressione dove $\alpha = 0$ e $\beta = 1$.

Che cosa si può concludere dai risultati di questa simulazione? Possiamo dire che, in base alla CTT,

- c'è una relazione lineare tra il punteggio osservato X e il punteggio vero T; tale relazione lineare ha pendenza unitaria e intercetta zero.
- La CTT fa proprie le assunzioni del modello di regressione lineare: incorrelazione tra variabile esplicativa T ed errore E, e indipendenza e gaussianità degli errori.
- Come conseguenza di tali assunzioni, il coefficiente di attendibilità non è altro che la quota di varianza del punteggio osservato X che viene spiegata dal punteggio vero tramite una regressione lineare, ovvero non è altro che il coefficiente di determinazione del modello di regressione $X = \alpha + \beta T + E$, dove $\alpha = 0$ e $\beta = 1$.

Vedremo in seguito come sia possibile formulare la CTT nei termini del modello statistico dell'analisi fattoriale. Nel linguaggio dell'analisi fattoriale, la varianza dell'errore σ^2_E viene chiamata *specificità* (*uniqueness*).

2.8 Misurazioni parallele e affidabilità

L'equazione $\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X}$ definisce il coefficiente di attendibilità ma non fornisce gli strumenti per calcolarlo, dato che la varianza del punteggio vero σ^2_T è una quantità incognita. Il metodo utilizzato dalla CTT per ottenere una stima (empirica) dell'attendibilità è quello delle forme parallele del test. Se è possibile elaborare versioni alternative dello stesso test che risultino equivalenti tra loro in termini di contenuto, modalità di risposta e caratteristiche statistiche, allora diventa anche possibile stimare il coefficiente di attendibilità.

Secondo la CTT, due test $X = T + E$ e $X' = T' + E'$ si dicono misurazioni parallele della stessa abilità latente se il punteggio vero T è uguale al punteggio vero T' e se la varianza degli errori $\mathbb{V}(E)$ è uguale alla varianza degli errori $\mathbb{V}(E')$.

Se il punteggio vero è uguale al valore atteso del punteggio osservato, $T = \mathbb{E}(X)$, allora devono essere uguali anche le medie dei punteggi osservati delle due forme parallele del test, $\mathbb{E}(X) = \mathbb{E}(X')$.

Dimostrazione. Consideriamo l'eguaglianza dei valori attesi dei punteggi osservati in due forme parallele del test: $\mathbb{E}(X) = \mathbb{E}(X')$. Risulta immediato che

$$\mathbb{E}(X) = \mathbb{E}(T + E) = \mathbb{E}(T) + \mathbb{E}(E) = T,$$

dato che $\mathbb{E}(E) = 0$ e T non è una variabile aleatoria. Inoltre, $\mathbb{E}(X') = T$, dato che $T = T'$. Ne segue che $\mathbb{E}(X) = \mathbb{E}(X')$. □

In maniera corrispondente, anche le varianze dei punteggi osservati di due misurazioni parallele devono essere uguali, $\mathbb{V}(X) = \mathbb{V}(X')$.

Dimostrazione. Per la misurazione parallela X abbiamo

$$\mathbb{V}(X) = \mathbb{V}(T + E) = \mathbb{V}(T) + \mathbb{V}(E);$$

per la misurazione parallela X' abbiamo

$$\mathbb{V}(X') = \mathbb{V}(T' + E') = \mathbb{V}(T') + \mathbb{V}(E').$$

Dato che $\mathbb{V}(E) = \mathbb{V}(E')$ e che $T = T'$, ne segue che $\mathbb{V}(X) = \mathbb{V}(X')$. □

Per costruzione, inoltre, gli errori E e E' devono essere incorrelati con T e tra loro.

2.8.1 La correlazione tra misurazioni parallele

Un'ulteriore assunzione della CTT è la seguente. La CTT assume che, data una serie di misurazioni parallele $X_1, X_2, X_3, ...$ e un arbitrario test Z, si ha

$$\rho(X_1, X_2) = \rho(X_1, X_3) = \rho(X_2, X_3) = ...$$

e

$$\rho(X_1, Z) = \rho(X_2, Z) = \rho(X_3, Z) = ...$$

ovvero, tutte le misurazioni parallele sono correlate tra loro nella stessa misura e ciascuna misurazione parallela correla nella stessa misura con qualunque altro test.

L'assunzione precedente può essere espressa, in maniera equivalente, come segue. Si consideri la matrice di correlazioni calcolata su tutto il dominio degli item (ovvero, la matrice delle correlazioni tra ciascuna

coppia di item nel dominio del costrutto). La correlazione media di questa matrice quantifica la capacità media di ciascun item di rappresentare il costrutto. La CTT assume che la correlazione di ciascun item con ciascuno degli altri sia costante (ovvero, uguale per qualunque coppia di item). Detto in altri termini: secondo la CTT ciascun item rappresenta il costrutto nella stessa misura. Questa è un'assunzione molto forte che si riflette, come vedremo, nella formula del coefficiente α di Cronbach utilizzata per misurare l'attendibilità come consistenza interna. È un'assunzione molto forte che raramente viene soddisfatta in pratica.

Secondo la CTT, dunque, forme parallele del test devono avere lo stesso valore atteso e la stessa varianza. Inoltre, ciascuna forma parallela deve correlare nella stessa misura con qualunque altro test. In che modo si differenziano allora le forme parallele del test? L'unica differenza tra le forme parallele del test riguarda il punteggio osservato: a causa dell'errore di misurazione $X \neq X'$.

Il concetto di forme parallele del test è estremamente importante per la CTT perché attraverso tale nozione diventa possibile giungere ad una stima empirica dell'attendibilità. Prima di presentare questo ultimo passaggio algebrico è però necessario calcolare la correlazione tra due misurazioni parallele.

2.8.2 La correlazione tra due forme parallele del test

Secondo la CTT, la correlazione tra due misurazioni parallele è uguale al rapporto tra la varianza del punteggio vero e la varianza del punteggio osservato. Ricordiamo che la varianza del punteggio osservato è uguale nelle due forme parallele del test: $\mathbb{V}(X) = \mathbb{V}(X')$.

Dimostrazione. Assumendo, senza perdita di generalità, che $\mathbb{E}(X) = \mathbb{E}(X') = \mathbb{E}(T) = 0$, possiamo scrivere

$$\begin{aligned}
\rho_{XX'} &= \frac{\sigma(X, X')}{\sigma(X)\sigma(X')} \\
&= \frac{\mathbb{E}(XX')}{\sigma(X)\sigma(X')} \\
&= \frac{\mathbb{E}[(T + E)(T + E')]}{\sigma(X)\sigma(X')} \\
&= \frac{\mathbb{E}(T^2) + \mathbb{E}(TE') + \mathbb{E}(TE) + \mathbb{E}(EE')}{\sigma(X)\sigma(X')}.
\end{aligned}$$

Ma $\mathbb{E}(TE) = \mathbb{E}(TE') = \mathbb{E}(EE') = 0$; inoltre, $\sigma(X) = \sigma(X') = \sigma_X$. Dunque,

$$\rho_{XX'} = \frac{\mathbb{E}(T^2)}{\sigma_X \sigma_X} = \frac{\sigma_T^2}{\sigma_X^2}. \tag{2.3}$$

□

Si noti come la (2.3) e l'equazione che definisce il coefficiente di attendibilità, ovvero $\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}$, riportano tutte e due la stessa quantità a destra dell'uguale. Otteniamo così un importante risultato. Il coefficiente di attendibilità, ovvero il quadrato del coefficiente di correlazione tra il punteggio osservato e il punteggio vero, è uguale alla correlazione tra il valore osservato di due misurazioni parallele:

$$\rho_{XT}^2 = \rho_{XX'}. \tag{2.4}$$

Tale risultato è importante perché consente di esprimere la quantità inosservabile ρ_{XT}^2 nei termini della quantità $\rho_{XX'}$ che può essere calcolata sulla base del punteggio osservato. Quindi, la stima di ρ_{XT}^2 si riduce alla stima di $\rho_{XX'}^2$. Per questa ragione, la (2.4) è forse la formula più importante della CTT.

2.8.3 La correlazione tra punteggio osservato e punteggio vero

Consideriamo ora la correlazione tra punteggio osservato e punteggio vero. La (2.4) si può scrivere come

$$\rho_{XT} = \sqrt{\rho_{XX'}}.$$

Tale risultato si può interpretare dicendo che la correlazione tra punteggio osservato e punteggio vero è uguale alla radice quadrata del coefficiente di attendibilità.

2.8.4 I fattori che influenzano l'attendibilità

Considerando le precedenti tre equazioni

$$\rho_{XT}^2 = \rho_{XX'}, \quad \rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}, \quad \rho_{XT}^2 = 1 - \frac{\sigma_E^2}{\sigma_X^2},$$

possiamo dire che ci sono tre modi equivalenti per concludere che l'attendibilità di un test è alta. L'attendibilità di un test è alta

1. quando la correlazione tra misurazioni parallele è alta,
2. quando la varianza del punteggio vero è grande relativamente alla varianza del punteggio osservato,
3. quando la varianza dell'errore di misura è piccola relativamente alla varianza del punteggio osservato.

Tali considerazioni hanno importanti implicazioni per le scelte che devono guidare la costruzione di un test. Si consideri, in particolare, l'equazione $\rho^2_{XT} = \rho_{XX'}$. Se interpretiamo $\rho_{XX'}$ come la correlazione tra due item, allora tale equazione ci fornisce un criterio per la scelta degli item da includere in un test: dobbiamo includere nel test gli item che correlano maggiormente tra loro. In questo modo, infatti, l'attendibilità del test aumenterà perché gli item inclusi nel test risultano maggiormente correlati con il punteggio vero.

2.9 Metodi alternativi per la stima del coefficiente di attendibilità

Come si stima in pratica l'affidabilità? Un modo grossolano (e molto impreciso) consiste nel somministrare allo stesso gruppo di individui lo stesso test in due differenti momenti e di calcolare il coefficiente di correlazione dei punteggi totali (*test-retest reliability*). McDonald (2013) afferma che tale procedura può essere giustificata in due modi diversi. La prima giustificazione è basata sull'assunzione che il valore vero non varia tra le due somministrazioni del test. Se le cose stanno in questo modo, gli errori saranno indipendenti e la correlazione tra il punteggio osservato nelle due somministrazioni ci fornirà una stima di $\rho_{XX'}$. Il problema è che non disponiamo di nessuno strumento per distinguere questa situazione ideale dal caso in cui viene violata l'assunzione dell'invarianza del punteggio vero. Una seconda giustificazione del metodo test-retest ci porta a definire il punteggio vero di retest come la componente del punteggio osservato che non varia tra le due somministrazioni. Il tal senso, il coefficiente di attendibilità viene concepito come un coefficiente di stabilità temporale. In generale, maggiore è l'intervallo temporale tra le

due somministrazioni, minore sarà il valore del coefficiente di stabilità temporale. Uno dei problemi del metodo test-retest è che due somministrazioni successive di un test ci forniscono soltanto un sottoinsieme delle possibili informazioni che verrebbero raccolte da uno studio longitudinale che copre un periodo temporale maggiore. Se tale studio longitudinale venisse eseguito, potremmo trovare la funzione che descrive la variazione del punteggio osservato in funzione del tempo. In generale, tale funzione non può essere descritta da un singolo parametro. Resta aperta la domanda di quale sia relazione tra questa funzione e il coefficiente di attendibilità.

Se sono disponibili due forme parallele dello stesso test, l'affidabilità può essere calcolata mediante il coefficiente di correlazione dei punteggi totali dei due test (*parallel-forms reliability*), valendo l'uguaglianza $\rho_{XX'} = \rho^2_{XT}$. Anche questo metodo, come il metodo del test-retest, non è esente da errori.

Il metodo di stima più diffuso è quello conosciuto come Cronbach's alpha (*internal consistency reliability*) originariamente ricavato da Kuder and Richardson (1937) per item dicotomici e poi generalizzato da Cronbach (1951) per item a risposte ordinali di qualunque tipo. L'idea su cui si basa consiste nel fatto che ogni singolo item del test, se confrontato con tutti gli altri, può essere usato per stimarne l'affidabilità. L'analisi degli item viene utilizzata per ottenere una stima della consistenza interna del test e valuta la misura in cui gli item del test sono espressione dello stesso costrutto.

3

Dati mancanti

Raramente un ricercatore si trova nella situazione fortunata nella quale un'analisi statistica (CFA o altro) può essere condotta utilizzando un set di dati in cui tutte le variabili sono presenti per tutte le osservazioni: i dati mancanti sono la norma piuttosto che l'eccezione nella pratica della ricerca.

3.1 Tipologie di dati mancanti

I dati mancanti possono verificarsi per una serie di motivi. Ad esempio, i dati possono mancare per disegno dello studio ("mancanza pianificata"), come ad esempio nei progetti di ricerca in cui i partecipanti al campione vengono selezionati casualmente per completare vari sottoinsiemi della batteria di valutazione completa (a causa di considerazioni pratiche come vincoli di tempo). In tali condizioni, si presume che i dati mancanti si distribuiscano in un modo completamente casuale rispetto a tutte le altre variabili nello studio. In generale, i meccanismi che determinano la presenza di dati mancanti possono essere classificati in tre categorie:

1. *valori mancanti completamente casuali* (*Missing Completely At Random*, MCAR). La probabilità di dati mancanti su una variabile non è collegata né al valore mancante sulla variabile, né al valore di ogni altra variabile presente nella matrice dati che si sta analizzando;
2. *valori mancanti casuali* (*Missing At Random*, MAR). I valori mancanti sono indipendenti dal valore che viene a mancare, ma dipendono da altre variabili, cioè i dati sulla variabile sono mancanti per categorie di partecipanti che potrebbero essere identificati da valori su altre variabili;

3. *valori mancanti non ignorabili* (*Missing Not At Random*, MNAR). La mancanza di un dato può dipendere sia dal valore del dato stesso che dalle altre variabili. Per esempio, se si studia la salute mentale e le persone depresse riferiscono meno volentieri informazioni riguardanti il loro stato di salute, allora i dati non sono mancanti per caso.

3.2 La gestione dei dati mancanti

Il passo successivo dopo la definizione dei meccanismi è quello della gestione dei dati mancanti. Sostanzialmente le scelte possibili sono due: l'eliminazione dei casi o la sostituzione dei dati mancanti. Un metodo semplice, indicato solo nel caso in cui l'ammontare dei dati mancanti è limitato e questi sono mancanti completamente a caso (MCAR), è quello di cancellare i casi (*case deletion*).

I modi per eliminare i casi sono due: *listwise deletion* e *pairwise deletion*. Nel primo caso si elimina dal campione ogni caso che ha dati mancanti. Le analisi avverranno quindi solo sui casi che hanno valori validi per tutte le variabili in esame. Si ha una maggiore semplicità di trattazione, tuttavia non si utilizza tutta l'informazione osservata (si riduce la numerosità campionaria e, quindi, l'informazione). Il secondo metodo è la pairwise deletion, che utilizza tutti i casi che hanno i dati validi su due variabili volta per volta. In questo modo si riesce a massimizzare la numerosità del campione da utilizzare, ma si tratta comunque di un metodo che presenta dei problemi, per esempio il fatto che con questo approccio i parametri del modello saranno basati su differenti insiemi di dati, con differenti numerosità campionarie e differenti errori standard.

Quando i dati non sono MCAR è opportuno sostituirli con appropriate funzioni dei dati effettivamente osservati (*imputation*). Di seguito sono indicati alcuni metodi.

1. *Mean Imputation*. Il dato mancante viene sostituito con la media della variabile. Questo metodo, utilizzato troppo spesso per la sua semplicità, riducendo la variabilità dei dati, ha invece effetti importanti su molte analisi dei dati e generalmente dovrebbe essere evitato.

2. *Regression Imputation.* Si tratta di un approccio basato sulle informazioni disponibili per le altre variabili. Si stima una equazione di regressione lineare per ogni variabile utilizzando le altre come predittori. Questo metodo offre il vantaggio di poter utilizzare dei rapporti esistenti tra le variabili per effettuare le valutazioni dei dati mancanti; tuttavia esso è usato raramente, in quanto amplifica i rapporti di correllazione tra le variabili; quindi se le analisi si baseranno su regressioni, questo metodo è sconsigliato.
3. *Multiple Imputation.* La tecnica multiple imputation, applicabile in caso di MAR, prevede che un dato mancante su una variabile sia sostituito, sulla base dei dati esistenti anche sulle altre variabili, con un valore che però comprende anche una componente di errore ricavata dalla distribuzione dei residui della variabile.
4. *Expectation-Maximization.* Un altro approccio moderno del trattamento dei dati mancanti è l'applicazione dell'algoritmo Expectation Maximization (EM). La tecnica è quella di stimare i parametri sulla base dei dati osservati, e di stimare poi i dati mancanti sulla base di questi parametri (fase E). Poi i parametri vengono nuovamente stimati sulla base della nuova matrice di dati (fase M), e così via. Questo processo viene iterato fino a quando i valori stimati convergono. Tuttavia, una limitazione fondamentale dell'utilizzo dell'algoritmo EM per calcolare le matrici di input per le analisi CFA/SEM è che gli errori standard risultanti delle stime dei parametri non sono consistenti. Pertanto, gli intervalli di confidenza e i test di significatività possono essere compromessi.

3.2.1 Metodo Direct ML

Benché vengano talvolta usati, i metodi precedenti sono stati presentati solo per ragioni storiche. Nella pratica concreta è meglio usare il metodo *Direct ML*, conosciuto anche come "raw ML" o "full information ML" (FIML), in quanto è generalmente considerano come il metodo migliore per gestire i dati mancanti nella maggior parte delle applicazioni CFA e SEM. Direct ML è esente dai problemi associati all'utilizzo dell'algoritmo EM e produce stime consistenti sotto l'ipotesi di normalità multivariata per dati mancanti MAR.

Intuitivamente, l'approccio utilizza la relazione tra le variabili per dedurre quali siano i valori mancanti con maggiore probabilità. Ad esempio, se due variabili, X e Y, sono correlate positivamente, allora se, per alcuni i, X_i è il valore più alto nella variabile, è probabile che anche il valore mancante Y_i sia un valore alto. FIML utilizza queste informazioni senza procedere all'imputazione dei valori mancanti, ma invece basandosi sulle stime più verosimili dei parametri della popolazione, ovvero massimizzando direttamente la verosimiglianza del modello specificato. Sotto l'assunzione di normalità multivariata, la funzione di verosimiglianza diventa

$$L(\mu, \Sigma) = \prod_i f(y_i \mid \mu_i, \Sigma_i),$$

laddove y_i sono i dati, μ_i e Σ_i sono i parametri della popolazione se gli elementi mancanti in y_i vengono rimossi. Si cercano i valori μ e Σ che massimizzano la verosimiglianza.

In `lavaan` l'applicazione di tale metodo si ottiene specificando l'argomento `missing = "ml"`.

3.2.2 Un esempio concreto

Per applicare il metodo *direct ML*, Brown (2015) prende in esame i dati reali di un questionario (un singolo fattore, quattro item, una covarianza di errore) con dati mancanti (N = 650). Leggiamo i dati dell'esempio:

```
d <- readRDS(here::here("data", "brown_table_9_1.RDS"))
head(d)
#>   subject s1 s2 s3 s4
#> 1    5760  2  0  1 NA
#> 2    5761  3  3  3 NA
#> 3    5763  2  4  4 NA
#> 4    5761  2  0  0 NA
#> 5    5769  2  1  1 NA
#> 6    5771  4  3  3 NA
```

Il modello viene specificato come segue (seguiamo Brown, 2015):

```
model <- "
  esteem =~ s1 + s2 + s3 + s4
  s2 ~~ s4
"
```

Adattiamo il modello ai dati:

```
fit <- cfa(model, data = d, missing = "ml")
```

È possibile identificare le configurazioni di risposte agli item che contengono dati mancanti:

```
fit@Data@Mp[[1]]$npatterns
#> [1] 5
```

```
pats <- fit@Data@Mp[[1]]$pat * 1L
colnames(pats) <- fit@Data@ov.names[[1]]
print(pats)
#>      s1 s2 s3 s4
#> [1,]  1  1  1  1
#> [2,]  1  1  1  0
#> [3,]  0  1  1  1
#> [4,]  1  0  1  1
#> [5,]  1  1  0  1
```

Possiamo ora esaminare la copertura della covarianza nei dati, ovvero la proporzione di dati disponibili per ciascun indicatore e per ciascuna coppia di indicatori:

```
coverage <- fit@Data@Mp[[1]]$coverage
colnames(coverage) <- rownames(coverage) <- fit@Data@ov.names[[1]]
print(coverage)
#>        s1     s2     s3     s4
#> s1 0.9615 0.9231 0.9231 0.6692
#> s2 0.9231 0.9615 0.9231 0.6692
#> s3 0.9231 0.9231 0.9615 0.6692
#> s4 0.6692 0.6692 0.6692 0.7077
```

Ad esempio, consideriamo l'item `s1`; se moltiplichiamo la copertura di questo elemento per la numerosità campionaria

```
650 * 0.9615385
#> [1] 625
```

possiamo concludere che questa variabile contiene 25 osservazioni mancanti; e così via.

Procediamo poi come sempre per esaminare la soluzione ottentua.

```
effectsize::interpret(fit)
#>     Name    Value Interpretation
#> 1    GFI 0.999449   satisfactory
#> 2   AGFI 0.992292   satisfactory
#> 3    NFI 0.999193   satisfactory
#> 4   NNFI 0.998978   satisfactory
#> 5    CFI 0.999830   satisfactory
#> 6  RMSEA 0.020238   satisfactory
#> 7   SRMR 0.004853   satisfactory
#> 8    RFI 0.995155   satisfactory
#> 9   PNFI 0.166532           poor
#> 10   IFI 0.999830   satisfactory
```

```
standardizedSolution(fit)
#>       lhs op    rhs est.std    se      z pvalue
#> 1  esteem =~     s1   0.737 0.020 37.086      0
#> 2  esteem =~     s2   0.920 0.013 68.651      0
#> 3  esteem =~     s3   0.880 0.013 66.432      0
#> 4  esteem =~     s4   0.905 0.016 55.400      0
#> 5      s2 ~~     s4  -0.886 0.216 -4.109      0
#> 6      s1 ~~     s1   0.456 0.029 15.554      0
#> 7      s2 ~~     s2   0.153 0.025  6.190      0
#> 8      s3 ~~     s3   0.225 0.023  9.636      0
#> 9      s4 ~~     s4   0.182 0.030  6.151      0
#> 10 esteem ~~ esteem   1.000 0.000     NA     NA
#> 11     s1 ~1          2.375 0.078 30.610      0
#> 12     s2 ~1          1.881 0.066 28.592      0
```

```
#> 13     s3 ~1          1.584 0.059 26.781      0
#> 14     s4 ~1          1.850 0.071 26.048      0
#> 15 esteem ~1          0.000 0.000     NA     NA
#>    ci.lower ci.upper
#> 1     0.698    0.776
#> 2     0.894    0.947
#> 3     0.854    0.906
#> 4     0.873    0.937
#> 5    -1.309   -0.463
#> 6     0.399    0.514
#> 7     0.104    0.201
#> 8     0.179    0.271
#> 9     0.124    0.240
#> 10    1.000    1.000
#> 11    2.223    2.527
#> 12    1.752    2.010
#> 13    1.468    1.700
#> 14    1.710    1.989
#> 15    0.000    0.000
```

4

Dati non gaussiani e categoriali

4.1 Dati non gaussiani

Negli esempi precedenti di questa dispensa è stato utilizzato lo stimatore di massima verosimiglianza (ML). Molti dei modelli CFA e SEM riportati nella letteratura di ricerca applicata utilizzano stime di ML. Tuttavia, ML è appropriata solo per dati multivariati normali, a livello di scala a intervalli (cioè, quando la distribuzione congiunta delle variabili continue è distribuita normalmente). Quando i dati continui si discostano marcatamente dalla normalità (cioè, forti asimmetria o curtosi), o quando alcuni degli indicatori non sono a livello di cala a intervalli (cioè, dati binari, politomici o ordinali), allora dovrebbe essere utilizzato uno stimatore diverso dalla ML.

La ricerca ha dimostrato che ML è robusta nel caso di lievi deviazioni nella normalità. Tuttavia, quando la non normalità è più pronunciata, è necessario utilizzare uno stimatore diverso da ML per ottenere risultati statistici affidabili (vale a dire, statistiche accurate sulla bontà dell'adattamento ed errori standard delle stime dei parametri). ML è particolarmente sensibile ad una eccessiva curtosi.

Le conseguenze dell'uso del ML in condizioni di grave non normalità includono

- valori eccessivi della statistica χ^2 del modello;
- sottostima degli indici di bontà dell'adattamento come TLI e CFI;
- sottostima degli errori standard delle stime dei parametri.

Questi effetti deleteri sono esacerbati con la diminuzione della dimensione del campione.

I due stimatori più comunemente usati per dati continui non normali sono

- ML robusto,
- minimi quadrati ponderati.

L'uso di WLS non è, in generale, raccomandato, a meno che le dimensioni del campione non siano molto grandi. Al contrario, la ricerca ha dimostrato che il metodo ML robusto fornisce uno stimatore adeguato rispetto a diversi livelli di non normalità.

Esaminiamo qui un esempio che usa gli stessi dati utilizzati da Brown (2015) nelle tabelle 9.5 – 9.7.

```
d <- readRDS(here::here("data", "brown_table_9_5_data.RDS"))
head(d)
#>   x1 x2 x3 x4 x5
#> 1  0  0  0  0  0
#> 2  0  0  0  0  0
#> 3  0  0  0  0  0
#> 4  4  2  2  1  1
#> 5  1  0  1  6  0
#> 6  0  0  0  0  0
```

Le statistiche descrittive dei dati dell'esempio mostrano valori eccessivi di asimmetria e di curtosi.

```
psych::describe(d)
#>    vars   n mean   sd median trimmed mad min max range
#> x1    1 870 1.47 2.17      0    1.01   0   0   8     8
#> x2    2 870 0.82 1.60      0    0.42   0   0   8     8
#> x3    3 870 1.27 2.07      0    0.78   0   0   8     8
#> x4    4 870 1.03 1.93      0    0.54   0   0   8     8
#> x5    5 870 0.61 1.52      0    0.18   0   0   8     8
#>    skew kurtosis   se
#> x1 1.51     1.25 0.07
#> x2 2.40     5.67 0.05
#> x3 1.80     2.34 0.07
#> x4 2.16     3.98 0.07
#> x5 3.10     9.37 0.05
```

Definiamo un modello ad un fattore e, seguendo Brown (2015), aggiungiamo una correlazione residua tra gli indicatori `X1` e `X3`:

```
model <- "
  f1 =~ x1 + x2 + x3 + x4 + x5
  x1 ~~ x3
"
```

Procediamo alla stima dei parametri utilizzando uno stimatore di ML robusto. La sintassi `lavaan` è la seguente:

```
fit <- cfa(model, data = d, mimic = "MPLUS", estimator = "MLM")
```

Possiamo esaminare la soluzione ottenuta con i soliti metodi:

```
standardizedSolution(fit)
#>    lhs op rhs est.std    se      z pvalue ci.lower
#> 1   f1 =~  x1   0.753 0.030 25.226      0    0.695
#> 2   f1 =~  x2   0.718 0.035 20.495      0    0.649
#> 3   f1 =~  x3   0.845 0.022 38.183      0    0.801
#> 4   f1 =~  x4   0.779 0.031 25.377      0    0.719
#> 5   f1 =~  x5   0.806 0.027 29.651      0    0.753
#> 6   x1 ~~  x3   0.414 0.061  6.777      0    0.294
#> 7   x1 ~~  x1   0.433 0.045  9.619      0    0.344
#> 8   x2 ~~  x2   0.484 0.050  9.623      0    0.386
#> 9   x3 ~~  x3   0.287 0.037  7.674      0    0.213
#> 10  x4 ~~  x4   0.393 0.048  8.202      0    0.299
#> 11  x5 ~~  x5   0.350 0.044  7.987      0    0.264
#> 12  f1 ~~  f1   1.000 0.000     NA     NA    1.000
#> 13  x1 ~1       0.677 0.020 33.542      0    0.637
#> 14  x2 ~1       0.514 0.018 28.848      0    0.479
#> 15  x3 ~1       0.612 0.019 32.539      0    0.575
#> 16  x4 ~1       0.533 0.018 29.758      0    0.498
#> 17  x5 ~1       0.400 0.016 25.596      0    0.369
#> 18  f1 ~1       0.000 0.000     NA     NA    0.000
#>    ci.upper
#> 1     0.812
#> 2     0.787
#> 3     0.888
#> 4     0.840
#> 5     0.859
```

```
#> 6     0.534
#> 7     0.521
#> 8     0.583
#> 9     0.360
#> 10    0.486
#> 11    0.436
#> 12    1.000
#> 13    0.717
#> 14    0.549
#> 15    0.649
#> 16    0.568
#> 17    0.430
#> 18    0.000
```

```
effectsize::interpret(fit)
#>     Name   Value Interpretation
#> 1    GFI 0.99084   satisfactory
#> 2   AGFI 0.95420   satisfactory
#> 3    NFI 0.98975   satisfactory
#> 4   NNFI 0.97823   satisfactory
#> 5    CFI 0.99129   satisfactory
#> 6  RMSEA 0.07935           poor
#> 7   SRMR 0.01595   satisfactory
#> 8    RFI 0.97436   satisfactory
#> 9   PNFI 0.39590           poor
#> 10   IFI 0.99131   satisfactory
```

4.2 Dati categoriali

Quando almeno un indicatore è categoriale (cioè binario, politomico o ordinale), il metodo ML ordinario non dovrebbe essere utilizzato per stimare i modelli CFA. Le potenziali conseguenze del trattamento delle variabili categoriali come variabili continue in un'analisi CFA sono molteplici, incluso il fatto che può tale scelta può

- produrre stime attenuate delle relazioni tra indicatori, specialmente

quando ci sono effetti pavimento o soffitto;
- portare ad individuare "pseudofattori" che sono solo artefatti del metodo statistico;

- produrre distorsioni negli indici di bontà dell'adattamento e nella stima degli errori standard;
- produrre stime errate dei parametri.

Esistono vari stimatori che possono essere utilizzati con indicatori categoriali; ad esempio, minimi quadrati ponderati (WLS), minimi quadrati ponderati robusti (WLSMV) e minimi quadrati non ponderati (ULS).

4.2.1 Un esempio concreto

Nell'esempio discusso da Brown (2015), i ricercatori desiderano verificare un modello unifattoriale di dipendenza da alcol in un campione di 750 pazienti ambulatoriali. Gli indicatori di alcolismo sono item binari che riflettono la presenza/assenza di sei criteri diagnostici per l'alcolismo (0 = criterio non soddisfatto, 1 = criterio soddisfatto). I dati sono i seguenti:

```
d1 <- readRDS(here::here("data", "brown_table_9_9_data.RDS"))
head(d1)
#>   y1 y2 y3 y4 y5 y6
#> 1  1  1  1  1  1  1
#> 2  1  1  1  1  1  1
#> 3  1  1  1  1  1  0
#> 4  1  1  1  1  1  1
#> 5  0  0  0  0  0  0
#> 6  1  1  0  1  1  1
```

Il modello viene specificato nel modo seguente:

```
model1 <- "
  etoh =~ y1 + y2 + y3 + y4 + y5 + y6
"
```

Adattiamo il modello specificando che i dati sono a livello di scala ordinale:

```
fit1 <- cfa(
  model1,
  data = d1,
  ordered = names(d1),
  estimator = "WLSMVS",
  mimic = "mplus"
)
```

Esaminiamo la soluzione ottenuta:

```
summary(fit1, fit.measures = TRUE)
#> lavaan 0.6-10 ended normally after 16 iterations
#>
#>   Estimator                                         DWLS
#>   Optimization method                           NLMINB
#>   Number of model parameters                        12
#>
#>   Number of observations                           750
#>
#> Model Test User Model:
#>                                                   Standard      Robust
#>   Test Statistic                                     5.651       9.540
#>   Degrees of freedom                                     9           9
#>   P-value (Chi-square)                               0.774       0.389
#>   Scaling correction factor                                      0.592
#>         mean and variance adjusted correction (WLSMV)
#>
#> Model Test Baseline Model:
#>
#>   Test statistic                              1155.845     694.433
#>   Degrees of freedom                                15           9
#>   P-value                                        0.000       0.000
#>   Scaling correction factor                                  1.664
#>
#> User Model versus Baseline Model:
#>
#>   Comparative Fit Index (CFI)                    1.000       0.999
#>   Tucker-Lewis Index (TLI)                       1.005       0.999
```

```
#>
#>   Robust Comparative Fit Index (CFI)                            NA
#>   Robust Tucker-Lewis Index (TLI)                               NA
#>
#> Root Mean Square Error of Approximation:
#>
#>   RMSEA                                          0.000       0.009
#>   90 Percent confidence interval - lower         0.000       0.000
#>   90 Percent confidence interval - upper         0.028       0.051
#>   P-value RMSEA <= 0.05                          0.999       0.944
#>
#>   Robust RMSEA                                                 NA
#>   90 Percent confidence interval - lower                    0.000
#>   90 Percent confidence interval - upper                       NA
#>
#> Standardized Root Mean Square Residual:
#>
#>   SRMR                                           0.031       0.031
#>
#> Weighted Root Mean Square Residual:
#>
#>   WRMR                                           0.519       0.519
#>
#> Parameter Estimates:
#>
#>   Standard errors                             Robust.sem
#>   Information                                   Expected
#>   Information saturated (h1) model          Unstructured
#>
#> Latent Variables:
#>                    Estimate  Std.Err  z-value  P(>|z|)
#>   etoh =~
#>     y1                1.000
#>     y2                0.822    0.072   11.392    0.000
#>     y3                0.653    0.092    7.097    0.000
#>     y4                1.031    0.075   13.703    0.000
#>     y5                1.002    0.072   13.861    0.000
#>     y6                0.759    0.076   10.011    0.000
```

```
#>
#> Intercepts:
#>                    Estimate  Std.Err  z-value  P(>|z|)
#>    .y1                0.000
#>    .y2                0.000
#>    .y3                0.000
#>    .y4                0.000
#>    .y5                0.000
#>    .y6                0.000
#>     etoh              0.000
#>
#> Thresholds:
#>                    Estimate  Std.Err  z-value  P(>|z|)
#>     y1|t1            -0.759    0.051  -14.890    0.000
#>     y2|t1            -0.398    0.047   -8.437    0.000
#>     y3|t1            -1.244    0.061  -20.278    0.000
#>     y4|t1            -0.795    0.051  -15.436    0.000
#>     y5|t1            -0.384    0.047   -8.148    0.000
#>     y6|t1            -0.818    0.052  -15.775    0.000
#>
#> Variances:
#>                    Estimate  Std.Err  z-value  P(>|z|)
#>    .y1                0.399
#>    .y2                0.594
#>    .y3                0.744
#>    .y4                0.361
#>    .y5                0.397
#>    .y6                0.653
#>     etoh              0.601    0.063    9.596    0.000
#>
#> Scales y*:
#>                    Estimate  Std.Err  z-value  P(>|z|)
#>     y1                1.000
#>     y2                1.000
#>     y3                1.000
#>     y4                1.000
#>     y5                1.000
#>     y6                1.000
```

Bibliografia

Allen, M. J. and Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Gulliksen, H. (1961). Measurement of learning and mental abilities. Technical report, Princeton University.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4):443–477.

Kline, P. (2013). *Handbook of psychological testing*. Routledge.

Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160.

Lewis, C. (1986). Test theory and psychometrika: The past twenty-five years. *Psychometrika*, 51(1):11–22.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Mauro, R. (1990). Understanding LOVE (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, 108(2):314–329.

McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press.

Nunnally, J. C. (1994). *Psychometric theory*. McGraw-Hill.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Samejima, F. (1983). Constant information model on the dichotomous response level. In *New horizons in testing*, pages 287–308. Elsevier.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, 15:201–293.

Thurstone, L. L. (1947). Multiple-factor analysis; a development and expansion of the vectors of mind.