Estimating ability for two samples

William Revelle
David M. Condon
Northwestern University

Abstract

Using IRT to estimate ability is easy, but how accurate are the estimate and what about multiple samples?

Contents

Introduction

This is a brief diversion in looking at what happens when we create items using IRT models and then try to fit them with the same IRT models. This is meant merely to get our ideas organized. The questions to address include how well the `score.irt` function provides statistics that match those of the items generated, and what happens with various restriction of range corrections.

Creating the data

First, create the data. 10000 subjects with 9 dichotomous items. Then select two subsets of these data, one the top 16%, the other the bottom 16%. (This same exercise can be repeated with more items and the results are basically the same, although the reliabilities are higher.)

```
> set.seed(42)
> items <- sim.irt(nvar=9,n=10000)
> hi <- subset(items$items,items$theta>1)
> hi.df <- data.frame(truth=subset(items$theta,items$theta>1),observed=hi)
>  dim(hi.df)
[1] 1554   10
```

---

contact: William Revelle revelle@northwestern.edu
Draft version of October 18, 2012

```
> lo <- subset(items$items,items$theta < -1)
> lo.df <-  data.frame(truth=subset(items$theta,items$theta < -1),observed=lo)
> dim(lo.df)
[1] 1648   10
> describe(items$items)
   var     n mean   sd median trimmed mad min max range  skew kurtosis   se
V1   1 10000 0.93 0.26      1    1.00   0   0   1     1 -3.28     8.76 0.00
V2   2 10000 0.86 0.34      1    0.95   0   0   1     1 -2.11     2.46 0.00
V3   3 10000 0.78 0.42      1    0.84   0   0   1     1 -1.32    -0.26 0.00
V4   4 10000 0.65 0.48      1    0.69   0   0   1     1 -0.63    -1.60 0.00
V5   5 10000 0.50 0.50      1    0.50   0   0   1     1 -0.01    -2.00 0.01
V6   6 10000 0.35 0.48      0    0.31   0   0   1     1  0.63    -1.60 0.00
V7   7 10000 0.22 0.41      0    0.15   0   0   1     1  1.36    -0.16 0.00
V8   8 10000 0.13 0.33      0    0.04   0   0   1     1  2.23     2.95 0.00
V9   9 10000 0.07 0.25      0    0.00   0   0   1     1  3.45     9.91 0.00
> describe(hi.df)
            var    n mean   sd median trimmed  mad min  max range  skew kurtosis   se
truth         1 1554 1.54 0.45   1.42    1.47 0.39   1 4.33  3.33  1.38     2.50 0.01
observed.V1   2 1554 0.99 0.12   1.00    1.00 0.00   0 1.00  1.00 -8.03    62.50 0.00
observed.V2   3 1554 0.97 0.17   1.00    1.00 0.00   0 1.00  1.00 -5.68    30.30 0.00
observed.V3   4 1554 0.95 0.22   1.00    1.00 0.00   0 1.00  1.00 -4.00    13.99 0.01
observed.V4   5 1554 0.89 0.31   1.00    0.99 0.00   0 1.00  1.00 -2.49     4.20 0.01
observed.V5   6 1554 0.82 0.38   1.00    0.90 0.00   0 1.00  1.00 -1.67     0.78 0.01
observed.V6   7 1554 0.68 0.47   1.00    0.72 0.00   0 1.00  1.00 -0.77    -1.41 0.01
observed.V7   8 1554 0.49 0.50   0.00    0.48 0.00   0 1.00  1.00  0.05    -2.00 0.01
observed.V8   9 1554 0.34 0.47   0.00    0.30 0.00   0 1.00  1.00  0.67    -1.55 0.01
observed.V9  10 1554 0.18 0.39   0.00    0.11 0.00   0 1.00  1.00  1.62     0.64 0.01
> describe(lo.df)
            var    n  mean   sd median trimmed  mad   min max range  skew kurtosis   se
truth         1 1648 -1.53 0.44  -1.41   -1.47 0.41 -4.04  -1  3.04 -1.22     1.57 0.01
observed.V1   2 1648  0.80 0.40   1.00    0.88 0.00  0.00   1  1.00 -1.54     0.36 0.01
observed.V2   3 1648  0.66 0.47   1.00    0.70 0.00  0.00   1  1.00 -0.68    -1.54 0.01
observed.V3   4 1648  0.50 0.50   1.00    0.51 0.00  0.00   1  1.00 -0.02    -2.00 0.01
observed.V4   5 1648  0.34 0.48   0.00    0.31 0.00  0.00   1  1.00  0.66    -1.57 0.01
observed.V5   6 1648  0.21 0.41   0.00    0.13 0.00  0.00   1  1.00  1.45     0.09 0.01
observed.V6   7 1648  0.09 0.29   0.00    0.00 0.00  0.00   1  1.00  2.78     5.73 0.01
observed.V7   8 1648  0.05 0.23   0.00    0.00 0.00  0.00   1  1.00  3.94    13.55 0.01
observed.V8   9 1648  0.02 0.15   0.00    0.00 0.00  0.00   1  1.00  6.35    38.34 0.00
observed.V9  10 1648  0.01 0.12   0.00    0.00 0.00  0.00   1  1.00  8.28    66.58 0.00
```

Score using statistics from the entire set and then just from the reduced sets

Now do the scoring, first score all the items using all the subjects, compare the scores with "truth". The correlation of the estimate with truth is the square root of the reliability. Note that the reliability of this 9 item test is not very good, particularly at the extremes. For a 36 item test, the correlation was .91 (reliability = .82). Also note that the standard deviations of the fitted values do not match those of the generating model. Why is this?

```
> all <- irt.fa(items$items)


> all
```

```
Item Response Analysis using Factor Analysis

Call: irt.fa(x = items$items)
Item Response Analysis using Factor Analysis

 Summary information by factor and item
 Factor =  1
              -3    -2   -1    0    1     2     3
V1           0.12  0.15 0.15 0.11 0.06  0.03  0.02
V2           0.10  0.15 0.17 0.13 0.08  0.04  0.02
V3           0.09  0.17 0.22 0.19 0.11  0.05  0.02
V4           0.07  0.14 0.22 0.23 0.15  0.07  0.03
V5           0.04  0.11 0.21 0.26 0.21  0.11  0.04
V6           0.03  0.07 0.16 0.26 0.26  0.15  0.07
V7           0.02  0.05 0.11 0.18 0.22  0.17  0.09
V8           0.01  0.04 0.08 0.16 0.23  0.21  0.13
V9           0.02  0.03 0.06 0.10 0.14  0.14  0.11
Test Info    0.50  0.91 1.38 1.63 1.45  0.98  0.53
SEM          1.41  1.05 0.85 0.78 0.83  1.01  1.37
Reliability -0.99 -0.09 0.28 0.39 0.31 -0.02 -0.88

Factor analysis with Call: fa(r = r, nfactors = nfactors, n.obs = n.obs)

Test of the hypothesis that 1 factor is sufficient.
The degrees of freedom for the model is 27  and the objective function was  0.02
The number of observations was  10000  with Chi Square =  204.05  with prob <  4.2e-29

The root mean square of the residuals (RMSA) is  0.01
The df corrected root mean square of the residuals is  0.02

Tucker Lewis Index of factoring reliability =  0.977
RMSEA index =  0.026  and the 90 % confidence intervals are  0.022 0.029
BIC =  -44.63
> all.scores <- score.irt(all,items$items)
> all.df <- data.frame(truth=items$theta,estimate=all.scores$theta1)
> lowerCor(all.df)
         truth estmt
truth    1.00
estimate 0.73  1.00
> describe(all.df)
         var     n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
truth      1 10000 -0.01 1.01  -0.01   -0.01 1.00 -4.04 4.33  8.37  0.01    -0.01 0.01
estimate   2 10000 -0.02 1.67  -0.22   -0.02 1.74 -4.36 4.45  8.81 -0.02     0.23 0.02
```

This observed correlation between the underlying θ used to generate the data and the estimate of θ is a validity coefficient. Squaring this gives us the reliability.

Reliability = 0.53

Conventional estimates of reliability

We can find α and $\omega_{hierarchical}$ using the raw data.

```
> omega(items$items)
Omega
Call: omega(m = items$items)
```

```
Alpha:                 0.51
G.6:                   0.49
Omega Hierarchical:    0.44
Omega H asymptotic:    0.84
Omega Total            0.53

Schmid Leiman Factor loadings greater than  0.2
       g   F1*   F2*   F3*   h2   u2   p2
V1  0.21              0.23 0.10 0.90 0.45
V2  0.26                   0.09 0.91 0.74
V3  0.35        0.21       0.17 0.83 0.73
V4  0.36                   0.15 0.85 0.84
V5  0.37                   0.17 0.83 0.82
V6  0.37                   0.17 0.83 0.80
V7  0.32                   0.14 0.86 0.73
V8  0.27                   0.11 0.89 0.68
V9                         0.06 0.94 0.55

With eigenvalues of:
   g  F1*  F2*  F3*
0.85 0.16 0.07 0.08

general/max  5.38   max/min =   2.14
mean percent general =  0.7    with sd =  0.13 and cv of  0.18

The degrees of freedom are 12  and the fit is  0
The number of observations was  10000  with Chi Square =  10.17  with prob <  0.6
The root mean square of the residuals is  0
The df corrected root mean square of the residuals is  0.01
RMSEA index =  0  and the 90 % confidence intervals are  NA 0.009
BIC =  -100.35

Compare this with the adequacy of just a general factor and no group factors
The degrees of freedom for just the general factor are 27  and the fit is  0.01
The number of observations was  10000  with Chi Square =  99.76  with prob <  2.8e-10
The root mean square of the residuals is  0.01
The df corrected root mean square of the residuals is  0.02

RMSEA index =  0.016  and the 90 % confidence intervals are  0.013 0.02
BIC =  -148.92

Measures of factor score adequacy
                                                 g   F1*   F2*   F3*
Correlation of scores with factors             0.67  0.32  0.25  0.27
Multiple R square of scores with factors       0.45  0.10  0.06  0.07
Minimum correlation of factor score estimates -0.09 -0.79 -0.87 -0.85
>
```

Now do this with the tetrachoric correlations. Note that these values are over estimates of the correct values.

```
> omega(all$rho)
Omega
Call: omega(m = all$rho)
Alpha:                 0.72
G.6:                   0.7
Omega Hierarchical:    0.68
Omega H asymptotic:    0.88
```

```
Omega Total            0.76

Schmid Leiman Factor loadings greater than  0.2
      g   F1*   F2*   F3*   h2   u2   p2
V1 0.38              0.36 0.27 0.73 0.53
V2 0.42                   0.20 0.80 0.88
V3 0.50                   0.26 0.74 0.94
V4 0.48                   0.24 0.76 0.94
V5 0.49                   0.26 0.74 0.94
V6 0.51                   0.28 0.72 0.91
V7 0.50                   0.27 0.73 0.92
V8 0.46                   0.25 0.75 0.87
V9 0.39        0.92       1.00 0.00 0.15

With eigenvalues of:
   g  F1*  F2*  F3*
1.91 0.07 0.85 0.20

general/max  2.24   max/min =   13.05
mean percent general =  0.79    with sd =  0.27 and cv of  0.35

The degrees of freedom are 12  and the fit is  0

The root mean square of the residuals is  0.01
The df corrected root mean square of the residuals is  0.01

Compare this with the adequacy of just a general factor and no group factors
The degrees of freedom for just the general factor are 27  and the fit is  0.03

The root mean square of the residuals is  0.02
The df corrected root mean square of the residuals is  0.03

Measures of factor score adequacy
                                                 g   F1*  F2*   F3*
Correlation of scores with factors            0.83  0.18 0.97  0.42
Multiple R square of scores with factors      0.68  0.03 0.94  0.18
Minimum correlation of factor score estimates 0.37 -0.93 0.87 -0.65
```

Score based upon statistics from the reduced sets

Now, score the top and bottom scores using the stats from all the data.

```
> hi.scores <- score.irt(all,hi)
> describe(hi.scores)
       var    n mean   sd median trimmed  mad   min  max range  skew kurtosis   se
theta1   1 1554 1.86 1.27   1.50    1.81 1.54 -1.51 4.45  5.97  0.18    -0.20 0.03
total1   2 1554 0.70 0.13   0.67    0.70 0.16  0.33 1.00  0.67 -0.18    -0.04 0.00
fit1     3 1554 0.45 0.11   0.45    0.46 0.10  0.00 0.76  0.76 -1.30     3.38 0.00
> lo.scores <- score.irt(all,lo)
> describe(lo.scores)
       var    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
theta1   1 1648 -1.85 1.32  -1.51   -1.81 1.48 -4.36 1.50  5.85 -0.18    -0.49 0.03
total1   2 1648  0.30 0.13   0.33    0.30 0.16  0.00 0.67  0.67  0.10    -0.30 0.00
fit1     3 1648  0.44 0.12   0.44    0.45 0.10  0.00 0.72  0.72 -1.25     2.87 0.00
```

Now, do this again, but this time, find the stats for the hi and low groups separately.

```
> plot(all)
```

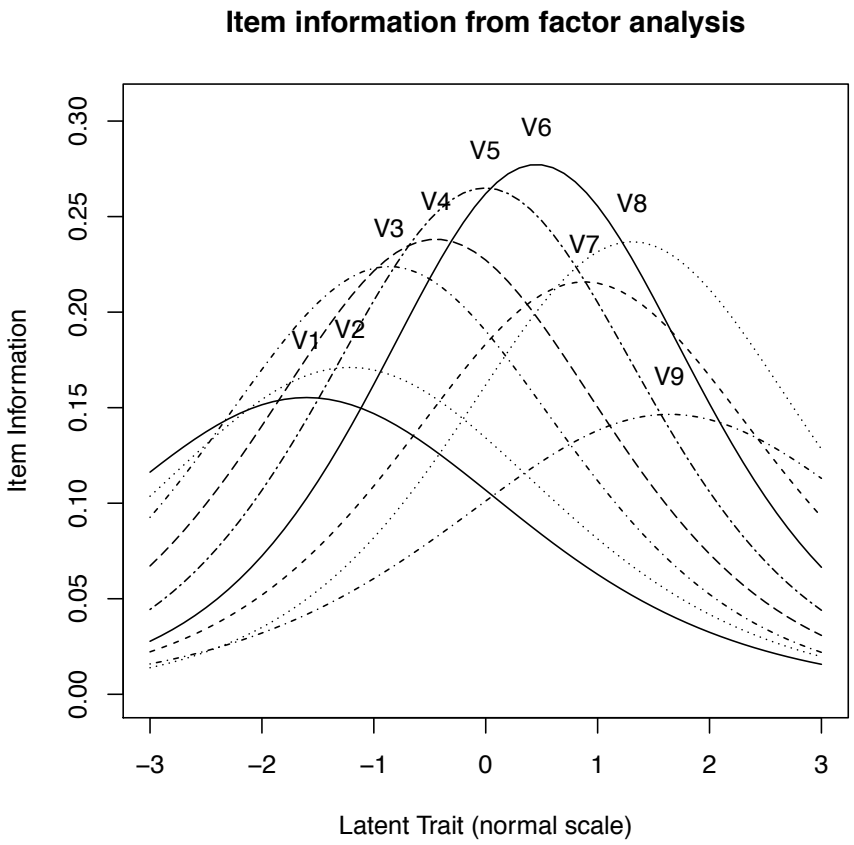

Figure 1. The item information curve for all the subjects.

```
> plot(all,type="test")
```

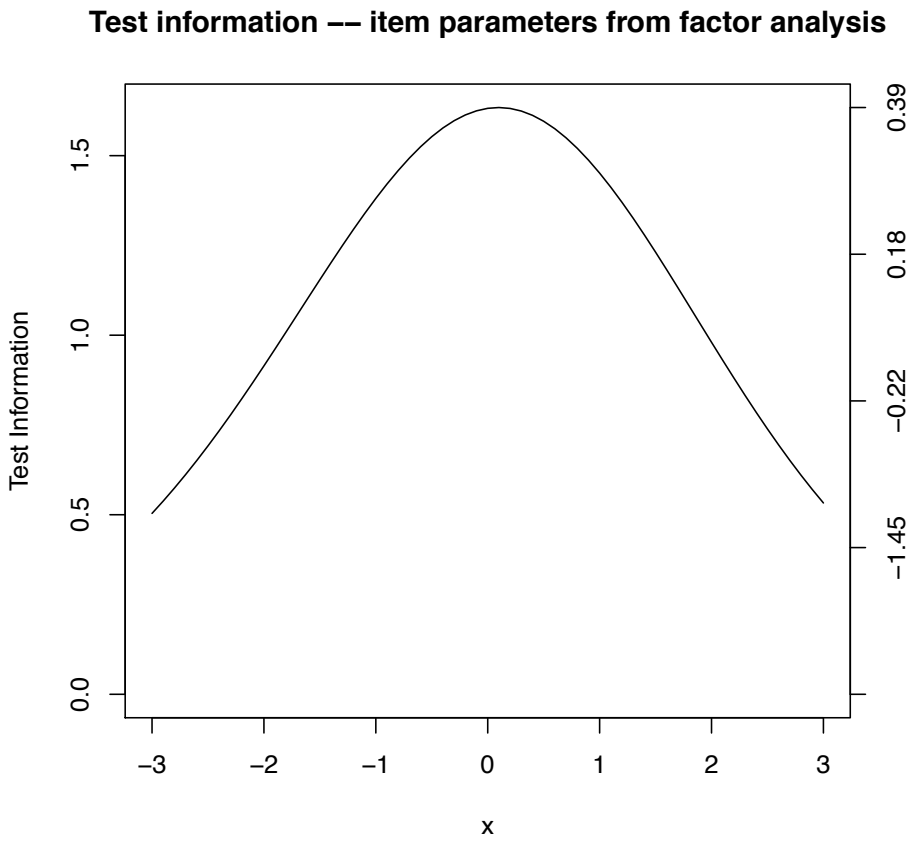

Figure 2. The test information curve suggests that it is not very good at the tails.

```
> hihi.stats <- irt.fa(hi)


> hihi.scores <- score.irt(hihi.stats,hi)
> describe(hihi.scores)

       var    n  mean   sd median trimmed mad  min  max range  skew kurtosis   se
theta1   1 1554  1.11 2.18   1.81    1.72   0 -5.0 5.00 10.00 -2.30     3.91 0.06
total1   2 1554 -0.05 0.17   0.00   -0.01   0 -0.5 0.50  1.00 -1.55     4.09 0.00
fit1     3 1554  0.02 0.05   0.00    0.00   0  0.0 0.15  0.15  2.48     4.17 0.00

> lowlow.stats <- irt.fa(lo)


> lowlow.scores <- score.irt(lowlow.stats,lo)
> describe(lowlow.scores)

       var    n  mean   sd median trimmed mad  min  max range skew kurtosis   se
theta1   1 1648 -0.22 0.56   -0.5   -0.32   0 -0.5 0.87  1.37 1.45     0.09 0.01
total1   2 1648  0.21 0.41    0.0    0.13   0  0.0 1.00  1.00 1.45     0.09 0.01
fit1     3 1648  0.00 0.00    0.0    0.00   0  0.0 0.00  0.00  NaN      NaN 0.00
```

Compare the means and correlations of these estimates

Compare the true values to the estimated values done from the IRT using the entire sample versus using the restricted sample. Notice that the standard deviations of the estimated scores are

```
> hi.est <- data.frame(truth=hi.df$truth,estimate=hi.scores$theta1,hiest=hihi.scores$theta)
> describe(hi.est)

         var    n mean   sd median trimmed  mad   min  max range  skew kurtosis   se
truth      1 1554 1.54 0.45   1.42    1.47 0.39  1.00 4.33  3.33  1.38     2.50 0.01
estimate   2 1554 1.86 1.27   1.50    1.81 1.54 -1.51 4.45  5.97  0.18    -0.20 0.03
hiest      3 1554 1.11 2.18   1.81    1.72 0.00 -5.00 5.00 10.00 -2.30     3.91 0.06

> lowerCor(hi.est)

         truth estmt hiest
truth    1.00
estimate 0.38  1.00
hiest    0.10  0.28  1.00

> lo.est <- data.frame(truth=lo.df$truth,estimate=lo.scores$theta1,loest=lowlow.scores$theta)
> describe(lo.est)

         var    n  mean   sd median trimmed  mad   min   max range  skew kurtosis   se
truth      1 1648 -1.53 0.44  -1.41   -1.47 0.41 -4.04 -1.00  3.04 -1.22     1.57 0.01
estimate   2 1648 -1.85 1.32  -1.51   -1.81 1.48 -4.36  1.50  5.85 -0.18    -0.49 0.03
loest      3 1648 -0.22 0.56  -0.50   -0.32 0.00 -0.50  0.87  1.37  1.45     0.09 0.01

> lowerCor(lo.est)

         truth estmt loest
truth    1.00
estimate 0.42  1.00
loest    0.19  0.44  1.00
```

Range Correction

One reason these correlations are smaller than the total group is restriction of range. Lets try a range correction based upon the standard deviations of the high (or low) scores adjusted for the population values.

```
> round(rangeCorrection(cor(hi.est)[1,2],sd(all.df$estimate),sd(hi.est$estimate)),2)
[1] 0.48
> round(rangeCorrection(cor(lo.est)[1,2],sd(all.df$estimate),sd(lo.est$estimate)),2)
[1] 0.51
```

Those estimates were based upon the concept that just the x variable was restricted (This is Thorndike Case 2). We can also do it on the basis of restriction the other way (case 1).

```
> round(rangeCorrection(cor(hi.est)[1,2],sd(all.df$estimate),sd(hi.est$estimate),case=1),2)
[1] 0.71
> round(rangeCorrection(cor(lo.est)[1,2],sd(all.df$estimate),sd(lo.est$estimate),case=1),2)
[1] 0.7
```

These values match the empirical estimates much more closely, but are we comfortable doing this correction this way?

This is a revision of the rangeCorrection function (not yet implemented into the *psych* package). The code for this is

```
rangeCorrection <-
function(r,sdu,sdr,sdxu=NULL,sdxr=NULL,case=2) {
if (!is.null(sdxu)) case <- 4  #
switch(case,
{ result <-  sqrt(1-(sdr^2/sdu^2) *(1-r^2))},
{ result <- (  r * sdu/(sdr* sqrt(1-r^2 + r^2*(sdu^2/sdr^2))))},
{result <- NULL},
{result <- r * (sdr/sdu)*(sdxr/sdxu) + (1-(sdr/sdu)^2) * (1- (sdxr/sdxu)^2 )  }
)
return(result)
}
```