Part 5: Testing the Rasch Model (I)

Goodness-of-Fit and some Diagnostics

Testing the RM – Overview

RM allows to evaluate the quality of measurement
crucial assumptions empirically testable
aim: find set of items that conform to the RM ('data fit model')

various classification of tests possible:
– according to assumptions of the RM
– types of statistics used (Pearson, LR, Wald)
– mathematical properties (distribution of test statistic)
– item/person oriented tests
– graphical procedures

we follow implementations in eRm

Andersen's Likelihood Ratio Test (Andersen, 1973)

– 'global' test (all items investigated simultaneously)
– powerful against violations of sufficiency and monotonicity
– can detect DIF (also called *item bias*):

Differential item functioning (DIF) occurs when individuals at the same level of an underlying trait differ in their responses to a specific item depending on certain characteristics (like age or gender)

basic idea:
consistent item parameter estimates ('invariance') can be obtained from a sample of any subgroup of population where the model holds

Andersen's Likelihood Ratio Test (cont'd)

divide the sample into $J-1$ groups
according to their total score r, $r = 1, \ldots, J-1$
obtain $J-1$ likelihoods of the form

$$L_c^{(r)} = \exp(-\sum_j \beta_j s_j^{(r)}) / \gamma(r; \beta_1, \ldots, \beta_J)^{n_r}$$

$s_j^{(r)}$ . . . number of correct responses to item j in scoregroup r.

the total likelihood is $L_c = \prod_r L_c^{(r)}$

then

$$\Lambda = \frac{L_c}{\prod_r L_c^{(r)}} = 1\,, \text{ only if the RM holds}$$

Andersen's Likelihood Ratio Test (cont'd)

Andersen (1973) proved that

$$Z = -2 \ln \Lambda = 2 \sum_r \ln L_c^{(r)} - 2 \ln L_c$$

is asymptotically χ^2-distributed with $df = (J-2)(J-1)$,
if $n_r \to \infty$ for all r.

practical problem:
– n often too small compared to number of different r's
way out:
– use ranges of r (2 or three subgroups)

test can be used for any partition of the sample according to extraneous variables (e.g., gender, age, ...)

using eRm: LRtest()

```
> load("stress.RData")
> rmod <- RM(stress)
> lr <- LRtest(rmod)
> lr

Andersen LR-test:
LR-value: 6.448
Chi-square df: 5
p-value:  0.265
```

factors as split criteria

```
> sex <- rep(c("male", "female"), each = 50)
> LRtest(rmod, splitcr = sex)
```

more than two groups (e.g., based on rawscores)

```
> gr3 <- cut(rowSums(stress), 3)
> LRtest(rmod, splitcr = gr3)
```


Wald Test

allows for testing single items

idea is again: sample into subgroups (usually 2)
as before, the item parameters should be invariant

for any partition of the sample into 2 groups:
using separate estimates $\hat{\beta}_j^{(1)}$ and $\hat{\beta}_j^{(2)}$ (and $\hat{\sigma}_{\beta_j}^{(1)}$, $\hat{\sigma}_{\beta_j}^{(2)}$,
we obtain

$$S_j = \frac{\hat{\beta}_j^{(1)} - \hat{\beta}_j^{(2)}}{\sqrt{\hat{\sigma}_\beta^{(1)} + \hat{\sigma}_\beta^{(2)}}} \approx N(0,1)$$

problem: statistics are not independent
but: allows for detecting DIF in single items

using eRm: Waldtest()

```
> wt <- Waldtest(rmod)
> wt

Wald test on item level (z-values):

        z-statistic p-value
beta I1      -0.045   0.964
beta I2       0.189   0.850
beta I3      -0.564   0.573
beta I4      -0.892   0.372
beta I5      -0.660   0.509
beta I6       2.463   0.014
```

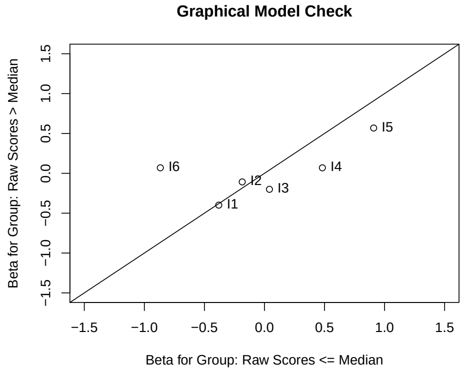
factors as split criteria (only for 2 levels)

```
> Waldtest(rmod, splitcr = sex)
```

both `LRtest()` and `Waldtest()` require objects obtained from `RM()`

Graphical Procedure

underlying idea again subgroup homogeneity, plot $\hat{\beta}^{(1)}$ vs $\hat{\beta}^{(2)}$



using eRm: `plotGOF()`

plot on last page:

```
> plotGOF(lr)
```

requires object obtained from `LRtest()`

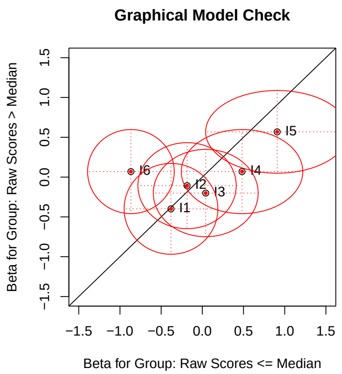
many options (see `?plotGOF`), two important are

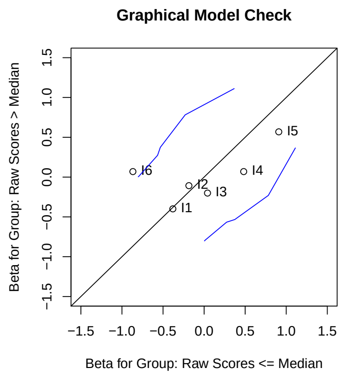
`conf=` . . . draws confidence ellipsoids
`ctrline=` . . . draws confidence bands

both options require `LRtest()` to be calculated using `se=TRUE`

```
> lr <- LRtest(rmod, se = T)
> plotGOF(lr, conf = list(), xlim = c(-1.5, 1.5), ylim = c(-1.5, 1.5))
> plotGOF(lr, ctrline = list(), xlim = c(-1.5, 1.5), ylim = c(-1.5, 1.5))
```


using eRm: `plotGOF()`





Person/Item Fit – Raw Data

objective is to detect noticeable patterns
– improbable patterns given a certain IRT model
– or too probable patterns (e.g., the deterministic Guttman Scale)

consider data

	I1	I2	I3	I4	I5	I6	sum
A	1	0	1	1	0	1	4
B	0	0	0	0	0	1	1
C	0	1	1	0	0	1	3
D	1	1	1	1	0	1	5
E	0	1	1	0	0	1	3
F	0	0	1	0	0	0	1
G	0	1	1	1	0	1	4
H	0	0	1	1	0	1	3
I	0	1	1	1	1	1	5
J	0	0	0	1	0	1	2
sum	2	5	8	6	1	9	–

Person/Item Fit - sorted data

noticeable patterns for persons

	I6	I3	I4	I2	I1	I5	r
D	1	1	1	1	1	0	5
I	1	1	1	1	0	1	5
A	1	1	1	0	1	0	4
G	1	1	1	1	0	0	4
C	1	1	0	1	0	0	3
E	1	1	0	1	0	0	3
H	1	1	1	0	0	0	3
J	1	0	1	0	0	0	2
B	1	0	0	0	0	0	1
F	0	1	0	0	0	0	1
s	9	8	6	5	2	1	–

noticeable patterns for items

	I6	I3	I4	I2	I1	I5	r
D	1	1	1	1	1	0	5
I	1	1	1	1	0	1	5
A	1	1	1	0	1	0	4
G	1	1	1	1	0	0	4
C	1	1	0	1	0	0	3
E	1	1	0	1	0	0	3
H	1	1	1	0	0	0	3
J	1	0	1	0	0	0	2
B	1	0	0	0	0	0	1
F	0	1	0	0	0	0	1
s	9	8	6	5	2	1	–

Person/Item Fit

Expected response: $\pi_{vi} = \exp(\theta_v - \beta_i)/(1 + \exp(\theta_v - \beta_i))$
Residuals: $e_{vi} = x_{vi} - \pi_{vi}$

Outfit MSQ: (unweighted mean-square)

for persons: $u_v = \frac{1}{k} \sum_i \frac{e_{vi}^2}{\pi_{vi}(1 - \pi_{vi})}$

for items: $u_i = \frac{1}{n} \sum_v \frac{e_{vi}^2}{\pi_{vi}(1 - \pi_{vi})}$

sensitive to unexpected rare extremes

Person/Item Fit (cont'd)

Infit MSQ: (information-weighted mean-square)

for persons: $w_v = \frac{1}{k} \frac{\sum_i e_{vi}^2}{\sum_i \pi_{vi}(1 - \pi_{vi})}$

for items: $w_i = \frac{1}{n} \frac{\sum_v e_{vi}^2}{\sum_v \pi_{vi}(1 - \pi_{vi})}$

sensitive to irregular inlying patterns

all have expectation 1, values $\neq 1$ indicate lack of fit
test statistics, e.g., nu_i^2, are χ^2 with corresponding df

Interpreting Infit/Outfit

Responses to Items:	Diagnosis of Pattern	Outfit	Infit
111 0110110100 000	Modelled/Ideal	1.0	1.1
111 1111100000 000	Guttman/Deterministic	0.3	0.5
000 0000011111 111	Miscode	12.6	4.3
011 1111110000 000	Carelessness/Sleeping	3.8	1.0
111 1111000000 001	Lucky Guessing	3.8	1.0
101 0101010101 010	Response set/Miskey	4.0	2.3
111 1000011110 000	Special knowledge	0.9	1.3
111 0101010101 000	Low discrimination	1.5	1.6
111 1110101000 000	High discrimination	0.5	0.7
111 1111010000 000	Very high discrimination	0.3	0.5

source: http://www.rasch.org/rmt/rmt82a.htm

items are arranged from easy to hard
vertical lines indicate zones where infit or outfit is more sensitive

Interpreting Infit/Outfit

Interpretation of parameter-level mean-square fit statistics: (rule of thumb)

>2.0	Distorts or degrades the measurement system
1.5 - 2.0	Unproductive for construction of measurement, but not degrading
0.5 - 1.5	Productive for measurement
<0.5	Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations

Person/Item Fit

Residuals, Infit and Outfit:

	I6	I3	I4	I2	I1	I5	Infit	Outfit
D	0.06	0.09	0.20	0.27	0.72	−0.93	0.58	0.25
I	0.06	0.09	0.20	0.27	−1.38	1.07	1.22	0.53
A	0.13	0.20	0.43	−1.74	1.53	−0.44	1.55	0.97
G	0.13	0.20	0.43	0.57	−0.65	−0.44	0.28	0.20
C	0.26	0.39	−1.20	1.11	−0.34	−0.23	0.85	0.51
E	0.26	0.39	−1.20	1.11	−0.34	−0.23	0.85	0.51
H	0.26	0.39	0.83	−0.90	−0.34	−0.23	0.51	0.32
J	0.48	−1.39	1.56	−0.48	−0.18	−0.12	1.28	0.81
B	0.96	−0.69	−0.32	−0.24	−0.09	−0.06	0.56	0.26
F	−1.04	1.45	−0.32	−0.24	−0.09	−0.06	1.20	0.56
Infit	0.63	0.96	0.99	0.99	0.97	0.60		
Outfit	0.25	0.50	0.67	0.71	0.56	0.26		

the blue areas indicate the noticeable patterns for items and persons, as already displayed in the sorted raw data

using eRm: `itemfit(),personfit()`

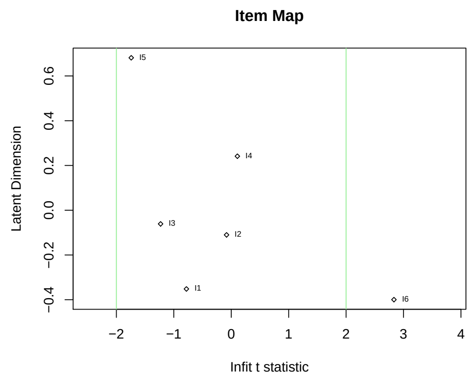
```
> personfit(pp)
> itemfit(pp)
```

both require objects obtained from `person.parameter()`

```
> pp <- person.parameter(rmod)
> itemfit(pp)
Itemfit Statistics:
     Chisq df p-value Outfit MSQ Infit MSQ Outfit t Infit t
I1  78.028 85   0.691      0.907     0.948    -1.12   -0.78
I2  86.465 85   0.435      1.005     0.992     0.11   -0.08
I3  77.243 85   0.713      0.898     0.918    -1.19   -1.23
I4  87.174 85   0.414      1.014     1.005     0.19    0.11
I5  73.947 85   0.798      0.860     0.863    -1.10   -1.74
I6 106.633 85   0.056      1.240     1.190     2.80    2.83
```


Graphical Procedure: `plotPWmap()`

underlying idea: plot item/personparameters vs. infit statistics



using eRm: `plotPWmap()`

plot on last page:

```
> plotPWmap(rmod)
```

can plot item and/or person locations
requires object obtained from `RM()`, or from `person.parameter()`

again many options (see `?plotPWmap`), some important are:

`imap=` . . . draws item map (default), subsets can be specified
`pmap=` . . . draws person map, subsets possible

`itemCI=` . . . confidence intervals for item locations
`personCI=` . . . confidence intervals for item locations

different colours, plotting symbols etc. possible


```
> plotPWmap(rmod, imap = TRUE, pmap = TRUE, person.subset = 1:5,
+     person.pch = 16, cex = 1,itemCI = list())
```