Bias reduction in the estimation of Rasch models

David Firth[1]
d.firth@warwick.ac.uk

Ioannis Kosmidis[21]
i.kosmidis@ucl.ac.uk

Heather Turner[1]
ht@heatherturner.net

[1]Department of Statistics,
University of Warwick

[2]Department of Statistical Science,
University College London

Psychoco, 2012

Outline

Rasch models

- Independent Bernoulli responses in a subject-item arrangement: Y_{is} is the outcome of the sth subject on the ith item.
- $\pi_{is} = P(Y_{is} = 1)$: the probability that sth subject succeeds on the ith item, $(i = 1, \ldots, I; s = 1, \ldots, S)$.

1PL model

- The 1PL Rasch model: (a special logistic regression model)

$$\log \frac{\pi_{is}}{1 - \pi_{is}} = \eta_{is} = \alpha_i + \gamma_s \quad (i = 1, \ldots, I; s = 1, \ldots, S) \,,$$

  where α_i, γ_s are uknown model parameters, and η_{is} the predictor for the 1PL model.

- Parameter vector: $\theta = (\alpha_1, \ldots, \alpha_I, \gamma_1, \ldots, \gamma_S)^T$,
- Parameter interpretation:
  - α_i (or $-\alpha_i$): measure of the "ease" (or "difficulty") of the ith item,
  - γ_s: the "ability" of the sth subject.

2PL model

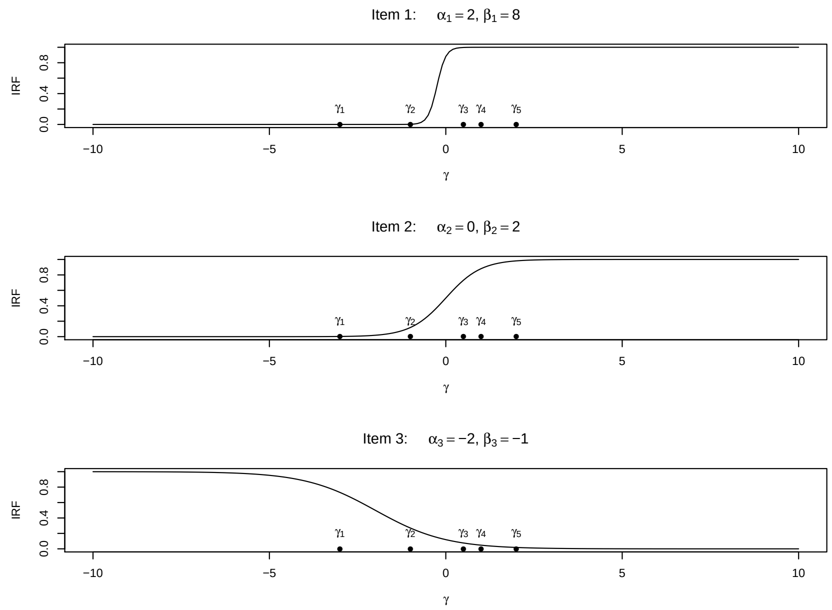
- The 2PL Rasch model:

$$\log \frac{\pi_{is}}{1-\pi_{is}} = \tilde{\eta}_{is} = \alpha_i + \beta_i \gamma_s \quad (i = 1, \ldots, I; s = 1, \ldots, S)\,,$$

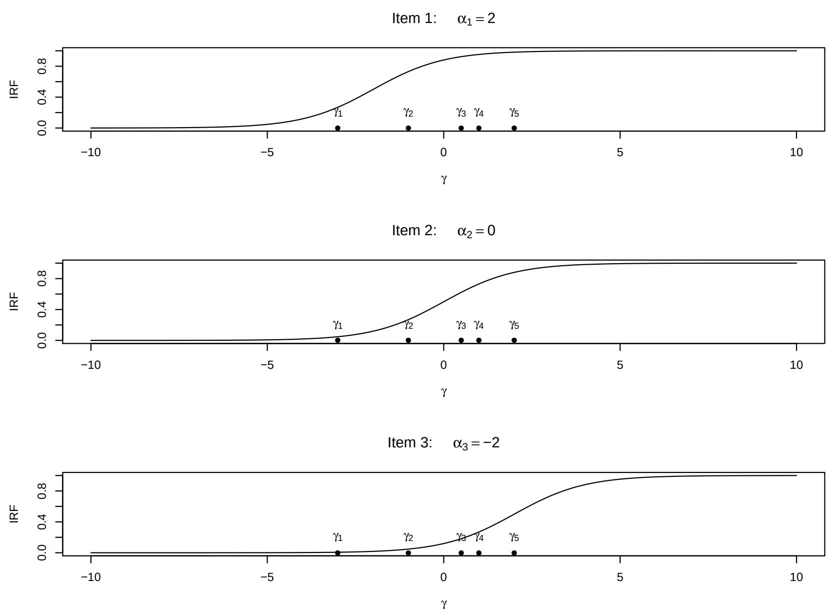
  where β_i is a "discrimination" parameter for the ith item, and $\tilde{\eta}_{is}$ the predictor for the 2PL model.

- Parameter vector: $\tilde{\theta} = (\alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_I, \gamma_1, \ldots, \gamma_S)^T$.
- The larger $|\beta_i|$ is the steeper is the Item-Response Function (IRF) (the map from γ_s to π_{is}).

2PL model: 5 subjects - 3 items



1PL model: 5 subjects - 3 items



Maximum likelihood estimation

→ ML estimation is straighforward using generic tools (e.g. `gnm` uses a quasi Newton-Raphon iteration).

→ Generic inferential procedures (LR tests, likelihood-based confidence intervals).

Maximum likelihood estimation - Issues

- Useful asymptotic frameworks (e.g. information grows with the number of subjects or number of items):
  → Full maximum likelihood generally delivers inconsistent estimates. (Andersen, 1980, Chapter 6)
  → Loss of performance (e.g. coverage) of tests, confidence intervals.

→ (Partial) Solutions: Conditional likelihoods, integrated likelihoods, modified profile likelihoods
  → can be hard to apply for 2PL due to nonlinearity.

Maximum likelihood estimation - Issues

- As with many models for binomial responses, there is positive probability of boundary ML estimates.
  → Numerical issues in estimation.
  → Problems with asymptotic inference (e.g. Wald-based inferences).

→ Add small constants to the responses in the spirit of Haldane (1955) (?)

Bias-reducing adjusted score functions

- Firth (1993): appropriate adjustment $A(\theta)$ to the score vector for getting estimators with smaller asymptotic bias than ML:

$$\nabla_\theta l(\theta) + A(\theta) = 0 \,.$$

- Applicable to models where the infromation on the parameters increases with the number of observations ($\dim \theta$ is independent of the number of observations).
  → Not the case for Rasch models under useful asymptotic frameworks.
  → But expect less-biased estimators than ML.

Bias-reducing adjusted score functions

→ In binomial/multinomial response GLMs, the reduced-bias estimates have been found to be always finite (Heinze and Schemper 2002; Bull et al. 2002; Zorn 2005; Kosmidis 2009)

→ Easy implementation:
  - Iterative bias correction (Kosmidis and Firth 2010)
  - Iterated ML fits on pseudo-data (Kosmidis and Firth 2011)

Adjusted score equations for 1PL

Adjusted score equations for 1PL (Firth 1993, logistic regressions)

$$0 = \sum_{i=1}^{I}\sum_{s=1}^{S}\left(y_{is} + \frac{1}{2}h_{is} + (1+h_{is})\pi_{is}\right) z_{ist} \quad (t = 1, \dots, I+S)\,,$$

where

- $z_{ist} = \partial\eta_{is}/\partial\theta_t$ is the (s,t)th element of the $S \times (I+S)$ matrix Z_i,
- h_{is} is the sth diagonal element of $H_i = Z_i F^{-1} Z_i^T \Sigma_r$ ("hat value" for the (i,s)th observation),
- $F = \sum_{i=1}^{T} Z_i^T \Sigma_i Z_i$ (the Fisher information),
- $\Sigma_i = \text{diag}\{v_{i1}, \dots, v_{iS}\}$, $v_{is} = \text{var}(Y_{is})$

Rasch Models | Maximum likelihood estimation | Bias reduction | Parameterization | Application | Discussion | References
Adjusted score functions

Adjusted score equations for 1PL

Adjusted score equations for 1PL (Firth 1993, logistic regressions)

$$0 = \sum_{i=1}^{I}\sum_{s=1}^{S}\left(y_{is} + \frac{1}{2}h_{is} + (1+h_{is})\pi_{is}\right) z_{ist} \quad (t = 1, \dots, I+S)\,,$$

where

- $z_{ist} = \partial\eta_{is}/\partial\theta_t$ is the (s,t)th element of the $S \times (I+S)$ matrix Z_i,
- h_{is} is the sth diagonal element of $H_i = Z_i F^{-1} Z_i^T \Sigma_r$ ("hat value" for the (i,s)th observation),
- $F = \sum_{i=1}^{T} Z_i^T \Sigma_i Z_i$ (the Fisher information),
- $\Sigma_i = \text{diag}\{v_{i1}, \dots, v_{iS}\}$, $v_{is} = \text{var}(Y_{is})$

Adjusted score equations for 2PL

Adjusted score equations for 2PL (Kosmidis and Firth 2009, GNMs)

$$0 = \sum_{i=1}^{I}\sum_{s=1}^{S}\left(y_{is} + \frac{1}{2}\tilde{h}_{is} + (1+\tilde{h}_{is})\pi_{is} + c_{is}v_{is}\right) \tilde{z}_{ist} \quad (t = 1, \dots, 2I+S)\,,$$

where

- $\tilde{z}_{ist} = \partial\tilde{\eta}_{is}/\partial\tilde{\theta}_t$ is the (s,t)th element of the $S \times (2I+S)$ matrix $\tilde{Z}_i$,
- $\tilde{h}_{is}$ is the "hat value" for the (i,s)th observation,
- $\tilde{F} = \sum_{i=1}^{T} \tilde{Z}_i^T \Sigma_i \tilde{Z}_i$,
- $\Sigma_i = \text{diag}\{v_{i1}, \dots, v_{iS}\}$, $v_{is} = \text{var}(Y_{is}) = \pi_{is}(1-\pi_{is})$,
- c_{is} is the asymptotic covariance of the ML estimators of β_i and γ_s (from the components of $\tilde{F}^{-1}$).

Rasch Models | Maximum likelihood estimation | Bias reduction | Parameterization | Application | Discussion | References
Adjusted score functions

Adjusted score equations for 2PL

Adjusted score equations for 2PL (Kosmidis and Firth 2009, GNMs)

$$0 = \sum_{i=1}^{I}\sum_{s=1}^{S}\left(y_{is} + \frac{1}{2}\tilde{h}_{is} + (1+\tilde{h}_{is})\pi_{is} + c_{is}v_{is}\right) \tilde{z}_{ist} \quad (t = 1, \dots, 2I+S)\,,$$

where

- $\tilde{z}_{ist} = \partial\tilde{\eta}_{is}/\partial\tilde{\theta}_t$ is the (s,t)th element of the $S \times (2I+S)$ matrix $\tilde{Z}_i$,
- $\tilde{h}_{is}$ is the "hat value" for the (i,s)th observation,
- $\tilde{F} = \sum_{i=1}^{T} \tilde{Z}_i^T \Sigma_i \tilde{Z}_i$,
- $\Sigma_i = \text{diag}\{v_{i1}, \dots, v_{iS}\}$, $v_{is} = \text{var}(Y_{is}) = \pi_{is}(1-\pi_{is})$,
- c_{is} is the asymptotic covariance of the ML estimators of β_i and γ_s (from the components of $\tilde{F}^{-1}$).

Pseudo data

→ If h and $\tilde{h}$ did not depend on the parameters then the reduced-bias estimator would be formally the ML estimator on Binomial pseudo-data.

Model	Pseudo-data
1PL	Responses: $y^* = y + h/2$ Totals: $m^* = 1 + h$
2PL	Responses: $\tilde{y}^* = y + \tilde{h}/2 + c\pi(1-\pi)$ Totals: $\tilde{m}^* = 1 + \tilde{h}$

Pseudo data

→ If h and $\tilde{h}$ did not depend on the parameters then the reduced-bias estimator would be formally the ML estimator on Binomial pseudo-data.

Model	Pseudo-data
1PL	Responses: $y^* = y + h/2$ Totals: $m^* = 1 + h$
2PL	Responses: $\tilde{y}^* = y + \tilde{h}/2 + c\pi 1_{(c>0)}$ Totals: $\tilde{m}^* = 1 + \tilde{h} + c(\pi - 1_{(c<0)})$

* via algebraic manipulation of the adjusted scores to ensure $0 \leq y^* \leq m^*$. Here, $1_E = 1$ if E holds.

Iterated ML fits on pseudo data

- The adjusted score equations can be solved as follows.

Iterated ML fits on pseudo data

At each iteration

1. Update the values of the pseudo data.
2. Use ML to fit the Rasch model on the current value of the pseudo data.

Repeat until the changes to the estimates are small.

- Ingredients: standard ML software, routines for extracting the hat values and Fisher information.

→ `gnm` and the methods `hatvalues`, `vcov` for `gnm` objects can do this

Iterated ML fits on pseudo data

- `tempFit`: a gnm object in identifiable parameterization, `pseudoData`: function that evaluates the pseudo data at the supplied fit — y^* and m^* depend on the parameters only through the "working weights" $\pi_{is}(1-\pi_{is})$.

```
## Rescale working weights:
tempFit$weights <- with(tempFit,  weights/prior.weights)
## Evaluate pseudo data
currentData <- pseudoData(tempFit)
## Fit model at the current pseudo data
tempFit <- update(tempFit, ys/ms ~ ., weights = ms,
                  data = currentData)
```

Identifiability

- 1PL model:

$$\log \frac{\pi_{is}}{1-\pi_{is}} = \eta_{is} = \alpha_i + \gamma_s \quad (i = 1, \ldots, I; s = 1, \ldots, S)\,,$$

- Fix location of α's or location of γ's (only $I + S - 1$ parameters can be estimated).
- Reduced-bias estimator is equivariant to ordinary constrasts (bias is equivariant in the group of affine transformations).

Identifiability

- 2PL model:

$$\log \frac{\pi_{is}}{1-\pi_{is}} = \eta_{is} = \alpha_i + \beta_i\gamma_s \quad (i = 1, \ldots, I; s = 1, \ldots, S)\,,$$

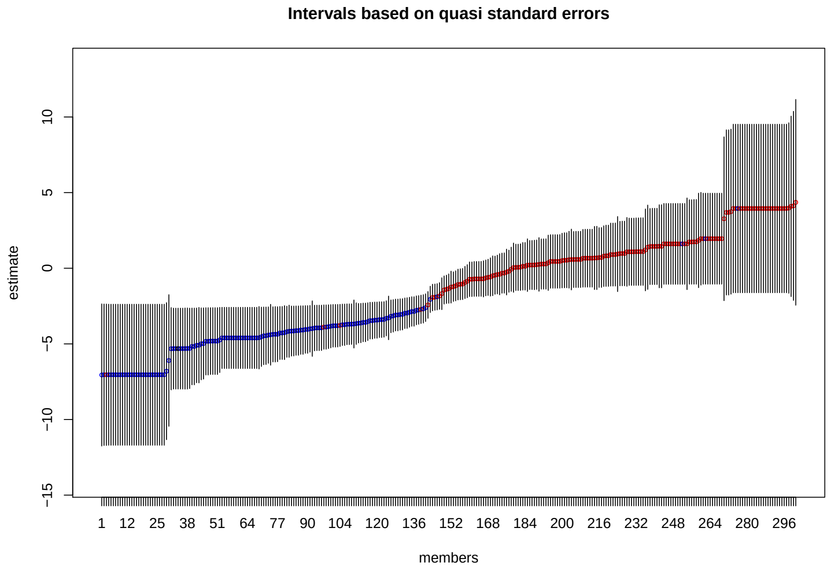
- Fix location of α's and scale of β's *or* location and scale of γ's (only $2I + S - 2$ parameters can be estimated).

Example: Scaling of legislators

Data:

- US House of Representatives, 20 roll calls selected by *Americans for Democratic Action*
- About 300 of 439 members voted on 10 or more of the 20 issues
- In `gnm` as dataset `House2001`; data kindly supplied by Jan deLeeuw, used in deLeeuw (2006, *CSDA*).
- Aim here is to place the members on a 'liberality' scale

`?House2001` in the `gnm` package uses an *ad hoc* (constant) data adjustment to achieve finite estimates for all 300 members. The method proposed in this talk is rather more principled!



This is very much **work in progress!**

The method described here yields more sensible results than either *MLE* or *constant* data-adjustment.

Computationally convenient.

But still it is *inconsistent* (e.g., as the number of items increases).

Aim of current work is to generalize fully the penalization approach of Firth (1993) to situations like this, where the number of 'nuisance' parameters increases with n.

Bull, S. B., C. Mak, and C. Greenwood (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis 39*, 57–74.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

Firth, D. and R. X. de Menezes (2004). Quasi-variances. *Biometrika 91*(1), 65–80.

Haldane, J. (1955). The estimation of the logarithm of a ratio of frequencies. *Annals of Human Genetics 20*, 309–311.

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine 21*, 2409–2419.

Kosmidis, I. (2009). On iterative adjustment of responses for the reduction of bias in binary regression models. Technical Report 09-36, CRiSM working paper series.

Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika 96*(4), 793–804.

Kosmidis, I. and D. Firth (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics 4*, 1097–1112.

Kosmidis, I. and D. Firth (2011). Multinomial logit bias reduction via the poisson log-linear model. *Biometrika 98*(3), 755–759.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis 13*, 157–170.