Item Response Theory in R: Models

Dr. Matthew Zeigenfuse

Lehrstuhl für Psychologisches Methodenlehre, Evaluation und Statistik
Psychologisches Institut
Universität Zürich

January 19, 2016

Part 1:
Preliminaries

Item Types

- **Dichtomous items**
  - Each person's response is either correct or incorrect
  - Incorrect responses are labelled 0 and correct responses are labelled 1
  - Any item type can be treated as a dichotomous item
- **Items with partial credit** (ordered polytomous items)
  - Each person's response is scored to one of K response categories labeled using the integers 0 to $K-1$ (shorthand: $0, \dots, K-1$)
  - Responses with higher scores are more correct
- **Multiple choice items** (nominal polytomous items)
  - Have K possible response categories labeled $0, \dots, K-1$
  - Non-response is often included as one of the K response categories

Data Matrix

- IRT models organize test takers' responses as a matrix
- Each element of the matrix contains the response of a single test taker to a single test item
- Each *row* of the matrix contains the responses of a single *test taker* to all of the test items
- Each *column* of the matrix contains the responses of all test takers to a single *test item*

Data Notation

- P is the number of test takers completing the test. The test takers are labelled $1, \ldots, P$
- I is the number of items the test contains. The test items are labelled $1, \ldots, I$
- The possible responses for item i are $0, \ldots, K_i - 1$
- $\mathbf{X}$ is the data matrix. It has P rows and I columns
- X_{pi} is the response of test taker p to test item i. It is an integer between 0 and $K_i - 1$
- $\mathbf{X}_{p\cdot}$ is row p, it contains the responses of participant p

Ability and Responding

- IRT models relate a test taker's ability to his or her pattern of responses
- A test taker's ability is a single number θ_p. It can be positive or negative
- Intuitively, test takers with higher ability tend to give higher responses
- This means test takers with high ability will correctly answer dichtomous items more often than those with low ability

Responses and Probability

- IRT models consider each test taker's responses to the items to be random
- The probability a test taker gives a particular response is determined by his or her ability
- A probability is a number between 0 and 1
- Responses with probabilities near 0 will rarely be offered and responses with probabilities near 1 will often be offered
- The probabilities of all available responses sum to one

Response Probabilities

- The probability that test taker p offers response k to item i given that his or her ability is θ_p is denoted

$$P(X_{pi} = k \mid \theta_p)$$

- This will be shortened to $P(X_{pi} \mid \theta_p)$ for expressions that do not depend on k
- The relationship between $P(X_{pi} = k \mid \theta_p)$ and θ_p is determined by the particular IRT model

Data Matrix Probability

- The probability of a data matrix can be computed from the $P(X_{pi} = 1 \mid \theta_p)$ by making two assumptions
  1. The test takers respond do not dependent on one another
  2. A given test taker's responses to the items of a test do not depend on one other
- Assumption 2 is known as local stochastic independence, or simply local independence
- Given these assumptions,

$$P(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{p=1}^{P} \prod_{i=1}^{I} P(X_{pi} \mid \theta_p)$$

- $\boldsymbol{\theta}$ is a vector whose elements are $\theta_1, \ldots, \theta_P$

Dichotomous vs. Polytomous Items

- Models for dichotomous items are the simplest IRT models
- These models are the basis for IRT models of polytomous items
- For this reason, we start with models for dichotomous items and work up to models for polytomous items

Part 2:
Models for Dichotomous Items

Specifying Dichotomous IRT Models

- Recall that the possible responses for a dichotomous item are 0 (incorrect) and 1 (correct)
- Since the sum of probabilities of these responses is one,

$$P(X_{pi} = 0 \mid \theta_p) = 1 - P(X_{pi} = 1 \mid \theta_p)$$

- Specifying $P(X_{pi} = 1 \mid \theta_p)$ as a function of θ_p defines a dichotomous IRT model
- This function is called the item response function and its graph is called the item characteristic curve

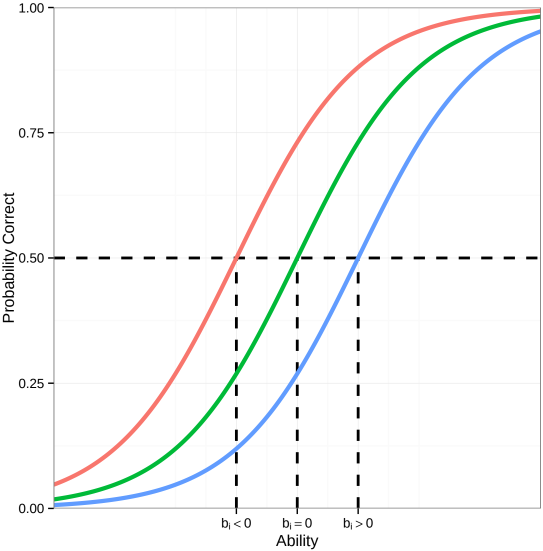
The Rasch Model

- The simplest IRT model is the Rasch model
- Its item response function is

$$P(X_{pi} = 1 \mid \theta_p) = \frac{1}{1 + e^{-(\theta_p - b_i)}}$$

- The item parameter b_i is known as the difficulty parameter
- It defines the ability for which the probability of correctly and incorrectly answering the item are both $1/2$
- The probability of correctly answering the item is
  - Greater than $1/2$ for abilities greater than b_i
  - Less than $1/2$ for abilities less than b_i

Rasch Model ICC



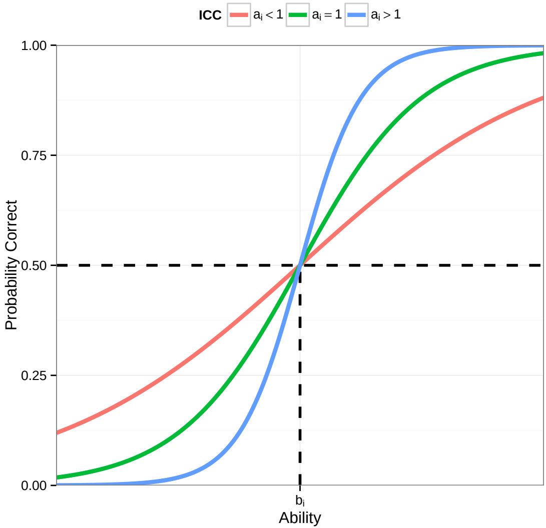
Birnbaum's Two-Parameter Logistic (2PL) Model

- Birnbuam's 2PL model augments the Rasch model with an additional parameter a_i controlling the slope of the ICC at $\theta = b_i$
- Its item response function is

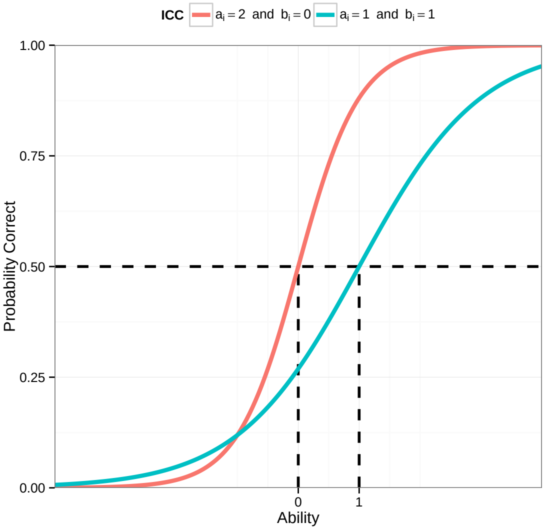
$$P(X_{pi} = 1 \mid \theta_p) = \frac{1}{1 + e^{-a_i(\theta_p - b_i)}}$$

- The parameter $a_i > 0$ is called the difficulty parameter
- $a_i = 1$ corresponds to the Rasch model

2PL ICC, Same Difficulty



2PL ICC, Different Difficulties



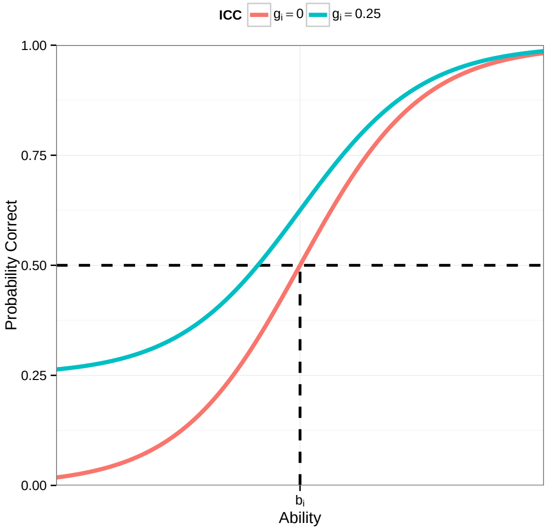
Guessing

- For many item types, test takers could correctly answer the item simply by guessing from the available responses
- Birnbaum's three-parameter logistic (3PL) accounts for this behavior by incorporating an additional parameter g_i controlling the lower asymptote of the item response function
- The ICC is

$$P(X_{pi} = 1 \mid \theta_p) = g_i + \frac{1 - g_i}{1 + e^{-a_i(\theta_p - b_i)}}$$

- g_i (approximately) defines the probability that a test taker with very low ability correctly responds to the item
- $g_i = 0$ corresponds to the 2PL model

3PL ICC


ICC
g_i = 0
g_i = 0.25
1.00
0.75
0.50
0.25
0.00
Probability Correct
b_i
Ability

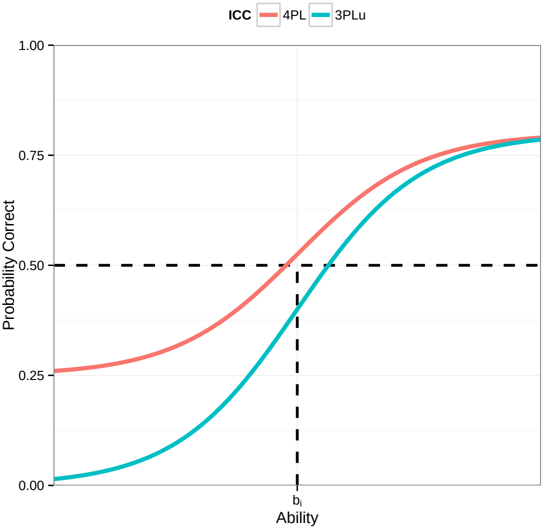
Slipping

- The 4PL model adds an additional parameter u_i which determines the upper asymptote of the item response function
- This allows test takers with very high ability to "slip" and incorrectly answer the item
- The item response function for the 4PL is

$$P(X_{pi} = 1 \mid \theta_p) = g_i + \frac{u_i - g_i}{1 + e^{-a_i(\theta_p - b_i)}}$$

- $u_i = 1$ corresponds to the 3PL
- `mirt` also provides another model, the 3PLu, with g_i fixed to 0 equation with

4PL ICC



ICC
4PL
3PLu
1.00
0.75
0.50
0.25
0.00
Probability Correct
b_i
Ability

Part 3:
Models for Partial Credit

Category Probability Functions

- The polytomous analogue to the item response function is the category probability function
- These functions relate ability to the probability of scoring in response category k
- For an item with K_i score categories, we need to specify $K_i - 1$ category response functions $P(X_{pi} = k \mid \theta_p)$
- Since probabilities sum to one, the function for category 0 is given by

$$P(X_{pi} = 0 \mid \theta_p) = 1 - \sum_{k=1}^{K_i} P(X_{pi} = k \mid \theta_p)$$

Partial Credit Model

- In the 2PL,

$$\log\left[\frac{P(X_{pi}=1 \mid \theta_p)}{P(X_{pi}=0 \mid \theta_p)}\right] = \alpha_i(\theta_p - \beta_i)$$

- The left-hand side is called the log-odds or logit
- Master's (1982) partial credit model extends the Rasch model by using it to model the logit between successive score categories, i.e.,

$$\ln\left[\frac{P(X_{pi}=k \mid \theta_p)}{P(X_{pi}=k-1 \mid \theta_p)}\right] = \theta_p - b_{i,k}$$

The Partial Credit Model

- The logit specification can be used to derive the category probability function

$$P(X_{pi} = k \mid \theta_p) \propto \begin{cases} 1, & \text{when } k = 0 \\ e^{k\theta_p - \sum_{m=1}^{k} b_{i,m}}, & \text{otherwise} \end{cases}$$

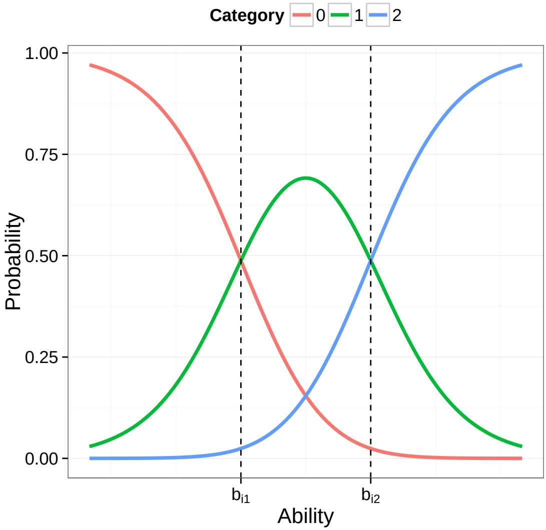
- The proportionality constant is

$$\left[1 + \sum_{n=1}^{K_i} e^{n\theta_p - \sum_{m=1}^{n} b_{i,m}}\right]^{-1}$$

Category Thresholds

- The $b_{i,k}$ are known as category thresholds
- Each category threshold $b_{i,k}$ determines where the category probability function for category k crosses that of $k - 1$
- They must be ordered such that $b_{i,1} < \ldots < b_{i,K_i - 1}$

PCM Category Probabilities



Generalized Partial Credit Model

- Muraki's (1992) generalized partial credit model (GPCM) augments the PCM with an discrimination parameter a_i, resulting in the model

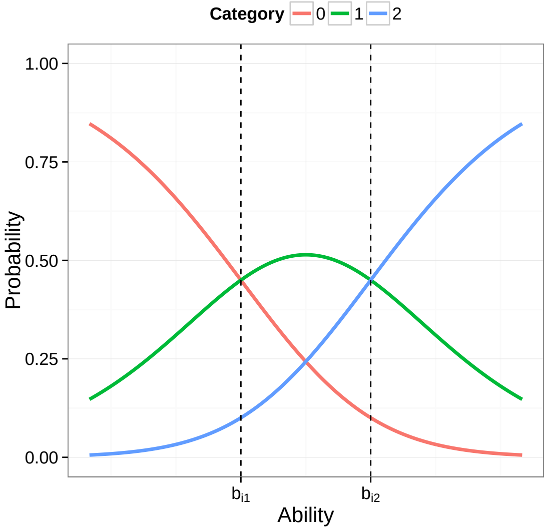
$$P(X_{pi} = k \mid \theta_p) \propto \begin{cases} 1, & \text{when } k = 0 \\ e^{a_i(k\theta_p - \sum_{m=1}^{k} b_{i,m})}, & \text{otherwise} \end{cases}$$

- The proportionality constant is

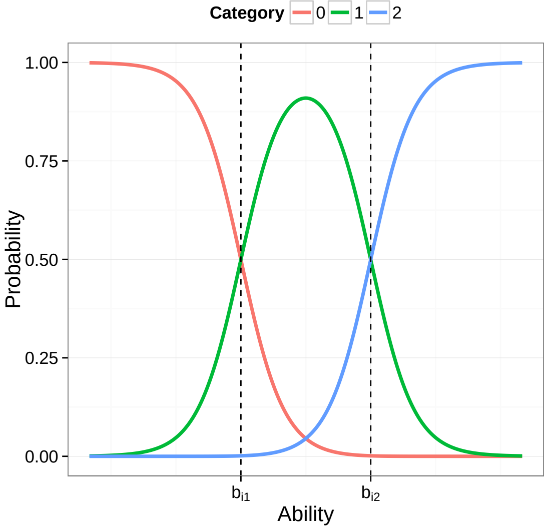
$$\left[1 + \sum_{n=1}^{K_i} e^{a_i(n\theta_p - \sum_{m=1}^{n} b_{i,m})}\right]^{-1}$$

- The discrimination parameter a_i controls the slope of the category probability function
- a_i corresponds to the PCM

GPCM Category Probabilities ($a_i = 1/2$)



GPCM Category Probabilities ($a_i = 2$)



Samejima's Graded Response Model (1)

- Samejima's (1969) graded response model does not extend the 2PL through the logit
- The probability of score k is the difference between the probabilities of scoring at least k and at least $k+1$, i.e.,

$$P(X_{pi} = k \mid \theta_p) = P(X_{pi} \geq k \mid \theta_p) - P(X_{pi} \geq k+1 \mid \theta_p)$$

- In

$$P(X_{pi} \geq 0 \mid \theta_p) = 1$$

and

$$P(X_{pi} \geq k \mid \theta_p) = \frac{1}{1 + e^{-a_{i,k} \cdot (\theta_p - b_{i,k})}}$$

for every $k \geq 1$

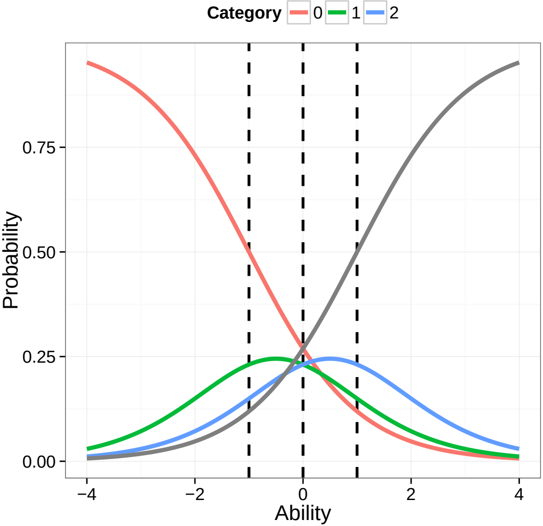
Samejima's Graded Response Model (2)

- This specification can be used to derive the following category probability functions

$$
\begin{aligned}
P(X_{pi} = 0 \mid \theta_p) &= \frac{1}{1 + e^{a_{i,1} \cdot (\theta_p - b_{i,1})}} \\
P(X_{pi} = k \mid \theta_p) &= \frac{e^{a_{i,k} \cdot (\theta_p - b_{i,k})} - e^{a_{i,k+1} \cdot (\theta_p - b_{i,k+1})}}{[1 + e^{a_{i,k} \cdot (\theta_p - b_{i,k})}][1 + e^{a_{i,k+1} \cdot (\theta_p - b_{i,k+1})}]} \\
P(X_{pi} = K_i - 1 \mid \theta_p) &= \frac{1}{1 + e^{-a_{i,K_i-1} \cdot (\theta_p - b_{i,K_i-1})}}
\end{aligned}
$$

- The center equality hold for $1 \leq k \leq K_i - 2$

Category Probability Curves



Graded Rating Scale Model

- The `mirt` package provides a simplification of the graded response model using "rating scale" parameterization of the $b_{i,k}$
- Andrich's (1978) rating scale model splits the threshold parameter into the sum of an item-specific part c_i and a category-specific part d_k
- This results in

$$b_{i,k} = c_i + d_k$$

- The category-specific part is shared among all items sharing the same rating scale

Part 4:
Models for Multiple Choice Items

Extending to Multiple Choice Items

- The 2PL can be extended to multiple choice items by
  1. Selecting a single response category as a reference
  2. Modeling the log-odds between the reference category and every other response category
- If we select 0 to be the reference category,

$$\log\left[\frac{P(X_{pi} = k \mid \theta_p)}{P(X_{pi} = 0 \mid \theta_p)}\right] = \alpha_{i,k} \cdot (\theta_p - \beta_{i,k})$$

- This indicates how much more likely the test taker is to select response category k versus the reference category

Bock's Nominal Categories Model

- We can use the fact that probabilities sum to one to derive a model that looks very similar to the GPCM

$$P(X_{pi} = k \mid \theta_p) \propto \begin{cases} 1, & \text{when } k = 0 \\ e^{\sum_{m=1}^{k} a_{i,m} \cdot (\theta_p - \beta_{i,m})}, & \text{otherwise} \end{cases}$$

- The proportionality constant is

$$\left[1 + \sum_{n=1}^{K_i} e^{\sum_{m=1}^{n} a_{i,m} \cdot (\theta_p - \beta_{i,m})}\right]^{-1}$$

- This is Bock's (1972) nominal categories models