**STAT 542: Multivariate Analysis** **Spring 2021**

# Lecture 6: Item response theory

*Instructor: Yen-Chi Chen*

The item response theory (IRT) is a popular model in psychometrics and educational statistics. In this note, we will briefly discuss the simplest IRT model called the *Rasch model*, which is named after Danish statistician Georg Rasch. A useful reference of this note is Chapter 12 of the following book:

> Sundberg, R. (2019). Statistical modelling by exponential families (Vol. 12). Cambridge University Press.

Suppose that we have $i = 1, \cdots, n$ individuals and $j = 1, \cdots, k$ items and the data is summarized as a $n \times k$ binary table/matrix $Y = \{Y_{ij}\}$, i.e., $Y_{ij} \in \{0, 1\}$. The Rasch model assumes that every cell (i.e., $Y_{ij}$) is independently drawn from a probability model:

$$P(Y_{ij} = 1; \alpha, \beta) = \frac{\alpha_i \beta_j}{1 + \alpha_i \beta_j}.$$

The parameter $\alpha_i \in \mathbb{R}$ denotes the $i$-th individual's ability and $\beta_j \in \mathbb{R}$ denotes the $j$-th item difficulty.

The goal is to infer the underlying parameters from the observed data table $Y$.

## 6.1 Joint Rasch Model

The Rasch model aims at finding both $\alpha$ and $\beta$ jointly. Since the Rasch model is essentially a collection of Bernoulli random variables, the joint PDF is

$$p(y; \alpha, \beta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \frac{(\alpha_i \beta_j)^{y_{ij}}}{1 + \alpha_i \beta_j},$$

which belongs to the exponential family.

Thus, after reparametrization, we obtain

$$\begin{aligned} p(y; \alpha, \beta) &\propto \prod_{i=1}^{n} \prod_{j=1}^{k} (\alpha_i \beta_j)^{y_{ij}} \\ &= \left( \prod_{i=1}^{n} \alpha_i^{\sum_{j=1}^{k} y_{ij}} \right) \left( \prod_{j=1}^{k} \beta_j^{\sum_{i=1}^{n} y_{ij}} \right) \\ &= \left( \prod_{i=1}^{n} \alpha_i^{y_{i+}} \right) \left( \prod_{j=1}^{k} \beta_j^{y_{+j}} \right), \end{aligned}$$

where

$$y_{i+} = \sum_{j=1}^{k} y_{ij}, \quad y_{+j} = \sum_{i=1}^{n} y_{ij}$$

are the row sum and column sum of the table $y$. Thus, the sufficient statistic of $\alpha_i$ is $Y_{i+}$ and the sufficient statistic of $\beta_j$ is $Y_{+j}$. The quantity $Y_{i+}$ is the number of 1 in $i$-th individual's response, which can be interpreted as the scores of $i$. The quantity $Y_{+j}$ is the number of 1 in item $j$, which can be interpreted as the number of individuals correctly answering question $j$.

Using the theory of exponential families, the MLE of $\alpha$ and $\beta$ can be obtained by solving the likelihood equations:

$$\begin{aligned} Y_{i+} &= \mathbb{E}(Y_{i+}; \widehat{\alpha}, \widehat{\beta}) = \sum_{j=1}^{k} \frac{\widehat{\alpha}_i \widehat{\beta}_j}{1 + \widehat{\alpha}_i \widehat{\beta}_j} \\ Y_{+j} &= \mathbb{E}(Y_{+j}; \widehat{\alpha}, \widehat{\beta}) = \sum_{i=1}^{n} \frac{\widehat{\alpha}_i \widehat{\beta}_j}{1 + \widehat{\alpha}_i \widehat{\beta}_j} \end{aligned} \tag{6.1}$$

for each $i = 1, \cdots, n$ and $j = 1, \cdots, k$.

Solving equation (5.1) is challenging so we need to do some reductions. First, there is an identifiability issue that if we multiply every $\alpha$ by a factor 10 and dividing all $\beta$ by the same factor, the probability model remains the same. So here we often set $\widehat{\beta}_1 = 1$ to avoid this problem. Second, each individual $Y_{i+} \in \{0, 1, 2, \cdots, k\}$ because there are only $k$ items. For individual with the same values of $Y_{i+}$, one would notice that their likelihood equations are identical! Namely, for each $s \in \{0, 1, 2, \cdots, k\}$ and $Y_{i+} = Y_{i'+} = s$,

$$\begin{aligned} s = Y_{i+} &= \sum_{j=1}^{k} \frac{\widehat{\alpha}_i \widehat{\beta}_j}{1 + \widehat{\alpha}_i \widehat{\beta}_j} \\ &= Y_{i'+} = \sum_{j=1}^{k} \frac{\widehat{\alpha}_{i'} \widehat{\beta}_j}{1 + \widehat{\alpha}_{i'} \widehat{\beta}_j}, \end{aligned}$$

which implies $\widehat{\alpha}_i = \widehat{\alpha}_{i'}$. Thus, $\alpha_i$ will only takes at most $k + 1$ distinct values. So now we reparametrize them as

$$\widehat{\theta}_1, \cdots, \widehat{\theta}_{k+1},$$

where

$$\widehat{\alpha}_i = \widehat{\theta}_s \text{ if } s = Y_{i+}.$$

Let

$$n_s = \sum_{i=1}^{n} I(Y_{i+} = s)$$

be the number of individuals with a score $s$. We can rewrite equation (5.1) as

$$\begin{aligned} s &= \sum_{j=1}^{k} \frac{\widehat{\theta}_s \widehat{\beta}_j}{1 + \widehat{\theta}_s \widehat{\beta}_j} = \widehat{\theta}_s \sum_{j=1}^{k} \frac{\widehat{\beta}_j}{1 + \widehat{\theta}_s \widehat{\beta}_j} \\ Y_{+j} = \mathbb{E}(Y_{+j}; \widehat{\alpha}, \widehat{\beta}) &= \sum_{s=0}^{k+1} n_s \frac{\widehat{\theta}_s \widehat{\beta}_j}{1 + \widehat{\theta}_s \widehat{\beta}_j} = \widehat{\beta}_j \sum_{s=0}^{k+1} n_s \frac{\widehat{\theta}_s}{1 + \widehat{\theta}_s \widehat{\beta}_j} \end{aligned} \tag{6.2}$$

for $s = 0, \cdots, k$ and $j = 2, \cdots, k$.

There are two special cases that we do not need to solve. The first one is $s = 0$ (all responses are 0); this leads to $\widehat{\theta}_0 = 0$. The other case is $s = k$, which leads to $\widehat{\theta}_k = \infty$. For other parameters, unfortunately, there is no closed-form solution to the above equations so we need other approach to find the estimators. Luckily,

the following iterative procedure provides a numerical solution to finding the MLE:

$$\begin{aligned}\widehat{\theta}_s^{(t+1)} &= s/\left(\sum_{j=1}^{k}\frac{\widehat{\beta}_j^{(t)}}{1+\widehat{\theta}_s^{(t)}\widehat{\beta}_j^{(t)}}\right)\\ \widehat{\beta}_j^{(t+1)} &= Y_{+j}/\left(\sum_{s=0}^{k+1} n_s\frac{\widehat{\theta}_s^{(t)}}{1+\widehat{\theta}_s^{(t)}\widehat{\beta}_j^{(t)}}\right)\end{aligned} \tag{6.3}$$

for $t = 0, 1, 2, \cdots$ with a proper initial parameter $\widehat{\theta}^{(0)}, \widehat{\beta}^{(0)}$.

Although we are able to compute the MLE, it may not be consistent! Namely, even if the model is correct, the MLE may not converge to the correct parameter values (for both individual-specific parameter $\alpha$ and item-specific parameter $\beta$). The major issue here is that the number of parameters increase linearly with respect to the sample size (number of individuals).

## 6.2 Conditional Rasch model

Here is an interesting property. Suppose that we are interested in only the item-specific parameters $\beta$, it can be consistently estimated using a modification called conditional Rasch model (consistent here refers to the number of individuals $n \to \infty$).

The idea of conditional Rasch model relies on asking the following question: suppose the individual $i$ correctly answers $s = Y_{i+}$ questions, what will the itemwise correct answers $\{Y_{i1}, \cdots, Y_{ik}\}$ informs us about the parameters $\beta_1, \cdots, \beta_k$. This question can be answered by the conditional distribution

$$p(y_{i1}, \cdots y_{ik}|Y_{i+} = s; \beta) \propto \prod_{j=1}^{k}\beta_j^{y_{ij}}.$$

The normalizing constant in the above probability will be

$$\sum_{y_{i1},\cdots y_{ik}:\sum_j y_{ij}=s}\prod_{j=1}^{k}\beta_j^{y_{ij}} = \gamma_s(\beta),$$

where $\gamma_s(\beta)$ is known as *elementary symmetric polynomial of degree s*. Note that

$$\gamma_0(\beta) = 0, \gamma_1(\beta) = \sum_j \beta_j, \quad \gamma_2(\beta) = \sum_{j_1<j_2}\beta_{j_1}\beta_{j_2}, \quad \gamma_3(\beta) = \sum_{j_1<j_2<j_3}\beta_{j_1}\beta_{j_2}\beta_{j_3}.$$

With this, we can rewrite

$$p(y_{i1}, \cdots y_{ik}|Y_{i+} = s; \beta) = \frac{\prod_{j=1}^{k}\beta_j^{y_{ij}}}{\gamma_s(\beta)}.$$

This defines the likelihood function of an individual $i$.

Using the fact that individuals are IID, the likelihood of using all individuals will be

$$\begin{aligned}L(\beta|Y) &= p(Y|Y_{1+}, \cdots, Y_{n+}; \beta)\\ &= \prod_{i=1}^{n}\frac{\prod_{j=1}^{k}\beta_j^{Y_{ij}}}{\gamma_{Y_{i+}}(\beta)}\\ &= \frac{\prod_{j=1}^{k}\beta_j^{Y_{+j}}}{\prod_{s=0}^{k}\gamma_s(\beta)^{n_s}},\end{aligned} \tag{6.4}$$

where $n_s = \sum_{i=1}^n I(Y_{i+} = s)$.

The MLE of $L(\beta|Y)$ solves the likelihood equations. Note that

$$\mathbb{E}(Y_{ij}|Y_{i+} = s) = \frac{\beta_i \cdot \gamma_{s-1}(\beta_{-j})}{\gamma_s(\beta)},$$

where $\beta_{-j} = \beta \backslash \{\beta_j\}$ is the vector of $\beta$ without $j$-th element. The likelihood equations will be

$$\begin{aligned} Y_{+j} = \sum_{i=1}^n Y_{ij} &= \sum_{i=1}^n \frac{\widehat{\beta}_j \cdot \gamma_{Y_{i+}-1}(\widehat{\beta}_{-j})}{\gamma_{Y_{i+}}(\widehat{\beta})} \\ &= \widehat{\beta}_j \cdot \sum_{s=0}^k n_s \frac{\gamma_{s-1}(\widehat{\beta}_{-j})}{\gamma_s(\widehat{\beta})}. \end{aligned}$$

Although solving the above equation is not simple, there has been several methods to it. In R, the package `eRm` is dedicated to the Rasch model and has a built-in function to solve this problem.

Unlike the joint Rasch model, the MLE in the conditional Rasch model leads to a consistent estimator of $\beta$. An intuitive way to understand this is that the single likelihood function (of $i$-th individual)

$$L(\beta|Y_i) = p(Y_{i1}, \cdots Y_{ik}|Y_{i+}; \beta) = \frac{\prod_{j=1}^k \beta_j^{Y_{ij}}}{\gamma_{Y_{i+}}(\beta)}$$

provides information on $\beta$. When sample size increases, the number of parameters remains fixed and the above likelihood function will be repeatedly used, leading to more and more information on the parameter. Thus, under regularity conditions, the MLE will be asymptotic normal and centered at $\beta$.

Here is an interesting note on the testing of conditional Rasch model. Since $Y_{i+} \in \{0, 1, 2, \cdots, k\}$, without using the Rasch model, we can characterize the distribution of $Y_i$ using

$$\beta_{js} : j = 1, \cdots, k; s = 0, 1, \cdots, k.$$

Namely, we allow the distribution to vary depending on the total number of correct answer. Note that by convention we set $\beta_{1s} = 1$ to avoid the identifiability issue (similar to the joint model). Also, similar to the joint model, there is no need to consider the case $s = 0, k$ since they corresponds to all 0's and all 1's. So the total free parameters without the Rasch model are

$$\beta_{js} : j = 2, \cdots, k; s = 1, \cdots, k-1.$$

Namely, there will be a total of $(k-1)(k-1)$ free parameters. With this, the Rasch model can be written as the null hypothesis

$$H_0 : \beta_{js} = \beta_j \text{ for all } s = 0, 1, \cdots, k; j = 1, \cdots k.$$

Under $H_0$ (Rasch model), there will be a total of $k-1$ free parameters. Thus, testing the feasibility of Rasch model is equivalent to testing $H_0$. There are a couple of ways we can perform this test such as the likelihood ratio test or the Wald test. Suppose that we use the likelihood ratio test, the test statistic will follows a $\chi^2$ distribution with a degree of freedom $(k-1)(k-1) - (k-2) = (k-2)(k-1)$.

## 6.3 Latent trait model

The latent trait model is a generalization of the Rasch model and is a popular model in item response theory. Similar to the conditional Rasch model, the latent trait model aims at recovering item-specific parameters.

The latent trait model uses a latent variable $Z_i$ for each individual that denotes the individual's ability. Let

$$\pi_j(Z_i) = P(Y_{ij} = 1|Z_i)$$

be a question-specific response function of the $j$-th question. The latent trait model assumes that

$$g(\pi_j(z)) = \xi_j + \eta_j z,$$

where $g$ is a known link function and $\xi_j, \eta_j$ are problem-specific parameters (later we will explain what they are). In a sense, the latent trait model is a generalized linear model.

A popular link function is the logit function, which corresponds to

$$\mathsf{logit}(\pi_j(z)) = \xi_j + \eta_j z.$$

With this, the probability model is

$$P(Y_{ij} = 1|Z_i; \xi, \eta) = \frac{\exp(\xi_j + \eta_j Z_i)}{1 + \exp(\xi_j + \eta_j Z_i)}.$$

This model reduces to the Rasch model when we choose $\eta_j = \eta_0$ for all $j$. The parameters $\alpha_i = \exp(\eta_0 Z_i)$ and $\beta_j = \exp(\xi_j)$.

Here is an interesting note. If we assume $Z_1, \cdots, Z_n \sim N(0, 1)$, then the logit model becomes

$$\mathsf{logit}(\pi_j(Z_i)) = \xi_j + \eta_j Z_i \sim N(\xi_j, \eta_j^2).$$

Thus, $\xi_j$ is often referred to as the *difficulty parameter* and $\eta_j$ is referred to as the *discrimination parameter*. If $\xi_j$ is very small (negatively large), then everyone has a low chance of correctly answering it. For the effect of $\eta_j$, for individual with high ability, i.e., $Z_i$ is large, $\eta_j Z_i$ will be large. So indeed $\eta_j$ describes how the individual's ability influences the chance of correctly answering item $j$.

When given the data table $Y$, we estimate $\widehat{\xi}, \widehat{\eta}$ via the ML procedure. This is a regular incomplete-likelihood problem so in general, there is no closed-form of the MLE. A common approach to numerically find the MLE is the EM algorithm[1]. In this problem, the $Q$ function in the EM algorithm (under the logit link function) will be

$$\begin{aligned} Q(\xi, \eta; \xi^{(t)}, \eta^{(t)}|Y_{ij}) &= \mathbb{E}\left(\log\left(\frac{\exp(\xi_j Y_{ij} + \eta_j Z_i Y_{ij})}{1 + \exp(\xi_j + \eta_j Z_i)}\right)\middle| Y_{ij}; \xi^{(t)}, \eta^{(t)}\right) \\ &= \xi_j Y_{ij} + \eta_j Y_{ij}\omega(Y_{ij}; \xi^{(t)}, \eta^{(t)}) - \underbrace{\mathbb{E}\left(\log\left(1 + \exp(\xi_j + \eta_j Z_i)\right)|Y_{ij}; \xi^{(t)}, \eta^{(t)}\right)}_{\Delta}, \\ \omega(Y_{ij}; \xi^{(t)}, \eta^{(t)}) &= \mathbb{E}(Z_i|Y_{ij}; \xi^{(t)}, \eta^{(t)}) = \frac{\int z \cdot \pi_j(z)^{Y_{ij}}(1 - \pi_j(z))^{1-Y_{ij}}\phi(z)dz}{\int \pi_j(z)^{Y_{ij}}(1 - \pi_j(z))^{1-Y_{ij}}\phi(z)dz}, \end{aligned}$$

where $\pi_j(z) = P(Y_{ij} = 1|Z_i = z; \xi, \eta) = \frac{\exp(\xi_j + \eta_j z)}{1 + \exp(\xi_j + \eta_j z)}$. In the E-step, we compute the conditional expectation of both $\omega$ and $\Delta$. In the Q-step, we attempt to find the maximizer. A numerical challenge here is that both $\omega$ and $\Delta$ do not have a simple closed-form. Thus, generally we use a Monte Carlo approximation to the expectation part; this is known as the Monte Carlo EM-algorithm.

[1] See http://faculty.washington.edu/yenchic/19A_stat535/Lec13_EM_SGD.pdf for an introduction