Provided for non-commercial research and educational use.
Not for reproduction, distribution or commercial use.

PLISKA
STUDIA MATHEMATICA
BULGARICA

ПЛИСКА
БЪЛГАРСКИ
МАТЕМАТИЧЕСКИ
СТУДИИ

The attached copy is furnished for non-commercial research and education use only.
Authors are permitted to post this version of the article to their personal websites or
institutional repositories and to share with other researchers in the form of electronic reprints.
Other uses, including reproduction and distribution, or selling or
licensing copies, or posting to third party websites are prohibited.

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **19** (2009), 59–68

PLISKA
STUDIA MATHEMATICA
BULGARICA

ESTIMATION OF IRT PARAMETERS OVER A SMALL SAMPLE. BOOTSTRAPPING OF THE ITEM RESPONSES

Dimitar Atanasov

Estimation of the parameters of of the Item Response Theory model is reasonable only on a relatively large samples. Applying this methodology for small samples is a common problem in practice. In this paper a bootstrapping technique for a small samples is presented. Additional item responses are added to the original dataset, according to the posterior probability of the correct item response. The same is used in generation of additional items needed when the cognitive attributes are studied.

1. Introduction

The aim of teaching process is to transfer abilities or knowledge from teachers to the students. As a result of that process students should perform some of these abilities or knowledge. A level of this performance has two origins. From one hand student's grade evaluate the the ability of students to recover the studied material. From the other hand, this could give a feedback for the way and methodology of theaching. Student abilities and recover of the knowledge can be be evaluated in many different ways, but may be one of the most frequently used in practice are different type of tests.

The test consists of set of questions (items) with closed (student should choose one answer from a given set) or open (student can write his own text) answers.

2000 *Mathematics Subject Classification:* 97C40.
Key words: item response theory, small sample, bootstraping.

Both of these types can be treated as test with dichotomous outcome 1 (the answer is correct) and 0 (the answer is wrong).

There are many theoretical constructs for modelling the result from a educational tests, but may of the most popular are Rasch model and its extension called Item Response Theory model (IRT) (Crocker & Algina 1986, Smith & Smith, 2004). According to the Rasch model, the probability of person n with ability θ_n succeeding on item i which has difficulty level D_i follows the equation

$$ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = \theta_n - D_i,$$

or equivalently

$$P_{ni} = \frac{\exp(\theta_n - D_i)}{1+\exp(\theta_n - D_i)}. \tag{1}$$

In 3-parametric IRT model, two additional parameters are included. The *discrimination* paramater a_i indexes how effectively the item discriminates between examinees who are relatively high on the criterion of interest and those who are relatively low. The *pseudo guessing* paramater c_i represents the probability that examinees with very low ability can guess the correct answer. Under this two additional parameters the probability for correct item response became

$$P_{ni} = c_i + (1-c_i)\frac{\exp(Ka_i(\theta_n - D_i))}{1+\exp(Ka_i(\theta_n - D_i))}, \tag{2}$$

where K is a constant which can be arbitrary set, but usually it is set to $K = 1.7$ because than P_{in} fits the normal ogive curve.

The probability for correct item response P_{ni} can be considered as a function $P_{ni}(\theta)$ of ability level of the examinees θ. Then, plotted against θ it gives so called Item Characteristic Curves (ICC). An example of ICC for two items are presented on Figure 1. The parameters of the *Item 2* are shown. The difficulty of the item is the ability level, giving probability or correct performance equal to 0.5 if there is no guessing. The discrimination of the item is presented by the slope of the tangent at the point of difficulty. The guess parameter represent the probability of correct item response (just by guessing) from subject with small level of abilities. In general, in this example, *Item 2* is more difficult, but less discriminative then *Item 1*, having larger value of the guessing parameter.

Estimation of the parameters of the test item can give important information for teachers. For example, which parts of the course (or which test items) are more difficult for students, how items cover the abilities under interest and so on. There are different techniques for estimating the item parameters (see for

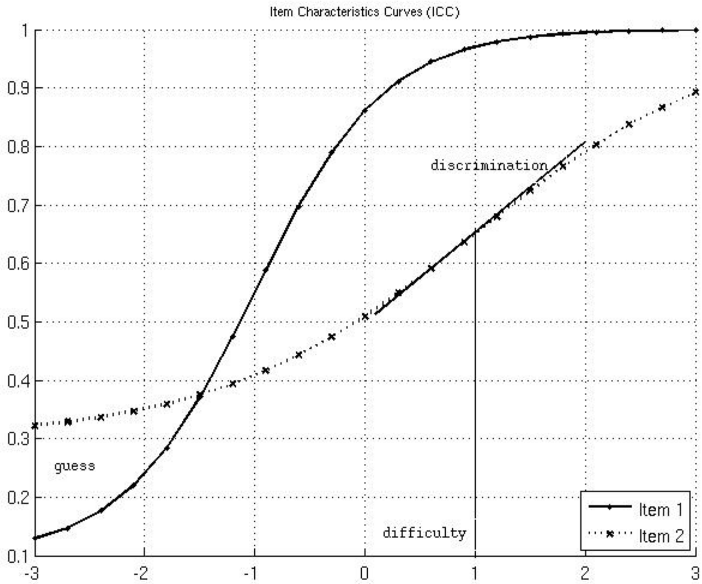

Figure 1: Item Characteristic Curves

example Smith & Smith, 2004) but may be the mots common method is to fit the empirical probabilities for correct item response with ICC Curves.

A different approach to the problem of student evaluation is to consider the cognitive attributes needed for correct item response. For example, an item from test in "Calculus" may require knowledge in *Integration* as well as knowledge in *Trigonometry.* So, the raw item responce is not sufficient if our aim is to estimate the level of knowledge only in *Integration.* To obtain such result one can use overall test performance and additional information for knowledge, required from any item, (so called attributes). Then having the ICC of the test items the performance of these attributes can be recovered.

Suppose that the set $C_1, \ldots, C_K$ represent the attributes. Now let us suppose that for a correct response on a given item, the student should possess all the attributes, needed by this item. Then (following Dimitrov, 2007), the probability for correct item response is given by

$$P_{ni}(\theta) = \prod_{k=1}^{K} \left(P(C_k = 1 \mid \theta)\right)^{q_{ik}}, \tag{3}$$

where $P(C_k = 1 \mid \theta)$ is the probability for correct performance on attribute k for a person with ability level θ and q_{ik} is 0/1 indicator that links item i to the attribute k. The matrix $Q = \{q_{ik}\}$ is so called *incidence Q-matrix*, with $q_{ik} = 1$ if item i requires attribute k and $q_{ik} = 0$ otherwise. Taking logarithm from both sides of the equation (3) Dimitrov (2007) obtain linear representation

$$\log P_{ni}(\theta) = \sum_{k=1}^{K} q_{ik} \log P(C_k = 1 \mid \theta), \tag{4}$$

or equivalently

$$L(\theta) = Q.X(\theta), \tag{5}$$

where $L(\theta)$ is known vector with elements $\log P_{ni}(\theta)$ and $X(\theta)$ is unknown vector with elements $P(C_k = 1 \mid \theta)$, representing the ICC of the attributes. Having probabilities of correct performance of the attributes one can eather recover the probabilities for correct response on a given item or fit the IRT parameters of an given attribute.

One of the main problems which arises using the IRT model for estimating the test item characteristics is that usually a large number (about 600) of observations (students) and a considerable large number of items (about 30) needed to obtain an estimation with a good properties (for example small standard error). In everyday practice it is allmost impossible to assure such large number of students which should pass given test. Even more, to estimate the performance on a set of relatively large number of cognitive attributes the test should contain large number of items.

There is no much information in the literature about the effect of the sample size to the properties of the estimated Rasch/IRT parameters. In practice, having a small sample, because of its simplicity, a Rasch model is preferable. A counterexample is given by de Gruijter (1986). The work of Stone & Yumoto (2004) shows that the sample size influence the estimates as might be expected and Rasch model gives the smallest goodness of fit index.

The effect of the sample size on the equating of the test items is considered by Ghada (2005), but the study is focused mainly to theproblem of calibration of the tests using samples with different sizes. An interesting approach about the critical relationship between item calibration, estimation of the IRT parameters and sample size in the context of Computer Adaptive Testing is performed by Ree & Jensen (1980). They show that except the guess parameter, the accuracy of of the estimation of the other two parameters strongly depends on the sample size and recommend: "*. . . the accurate estimate of the parameters requires large* *number of subjects over a broad range of ability ... therefore, it is necessary to administer test items, whether to be calibrated or equated to the largest samples available".*

In this paper two directions of bootstrapping are proposed. The first one is focused on estimation of IRT parameters of the items, included in the test, having relatively small number of students passing the test. The second direction gives additional set of artificial items in order to calculate the performance of the cognitive attributes.

2. Bootstrapping the item responses

Having relatively small number of observations, using the bootstrap technique one can generate a number of random sub samples of observations. The estimation of the parameters under interest over these sub samples can be treated as observations of a random variable. As a estimation of the value of these parameters a mean value of the estimated parameters over the sub samples will be used.

Let the test consist of p items $I_1, \dots, I_p$ and there are n examinees $X_1, \dots, X_n$.

Let $A_{n\times p} = (A_{ij})$, $i = 1, \dots, n$; $j = 1, \dots, p$ are the answers of the examinees. $A_{ij} = 1$ if examine i answers correctly to the item j and $A_{ij} = 0$ otherwise. Then the score S_i, $i = 1, \dots, n$ of the examine X_i is $\sum\limits_{k=1}^{p} A_{ik}$.

Let the examinees are grouped in r groups $G_1, \dots, G_r$, according to their score. $X_i \in G_k$ if $S_i \in (g_{k-1}, g_k)$, where $g_0, g_1, \dots, g_r$ is properly chosen set of score values.

To generate an answer of an artificial examine let us randomly choose a group G_c, which represents "knowledge" of the that examine X_b.

The probability of correct answer from examine X_b on item j is

$$P(A_j = 1 \mid X_b \in G_c) = \frac{P(X_b \in G_c \mid A_j = 1)P(A_j = 1)}{P(X_b \in G_c)}.$$

The probabilities in the right side of the equation can be replaced with empirical proportions

$$P(X_b \in G_c \mid A_j = 1) = \frac{\sharp\{X_i : X_i \in G_c, j \in \{j : A_{ij} = 1\}\}}{\sharp\{X_i : X_i \in G_c\}},$$

$$P(A_j = 1) = \frac{\sum\limits_{k=1}^{n} A_{ij}}{n},$$

$$P(X_b \in G_c) = \frac{\sharp\{X_i : X_i \in G_c\}}{n},$$

where with $\sharp$ the number of elements in the set is denoted.

Let us generate M sets of answers of N examinees ($N \gg n$). The generating algorithm can be summarized as follows (the algorithm can be referred as *resampling bootstrap method*, Chernick, 1999):

1. Setting starting values of the counters $m = 1$, $l = 1$, $h = 1$
2. Set $l = l + 1$ unless $l > N$,
3. Choose a random group G_c and set $m = m + 1$ unless $m > M$, otherwise set $m = 1$ and go back to 3
4. Set $h = h + 1$ unless $h > p$, otherwise set $h = 1$ and go back to 3
5. Flip a coin with probability $p = P(A_h = 1 \mid X_b \in G_c)$, $Bi(1, p)$, if it is head than set the answer as correct
6. Go back to 2

Over the generated M datasets the IRT parameters can be estimated. For example, let $D_{i1}, \ldots, D_{iM}$ are the estimated values of the difficulty parameter of the item i. Then, as a estimation of the difficulty parameter of the i-th item in the test the mean value

$$\hat{D}_i = \frac{\sum_{k=1}^{M} D_{ik}}{M}$$

can be used. The histogram of the estimated values $D_{i1}, \cdots, D_{iM}$ in the particular case of $n = 20$, $M = 1000$, $N = 300$ and $\hat{D}_i = -0.92$ is presented on Figure 2.

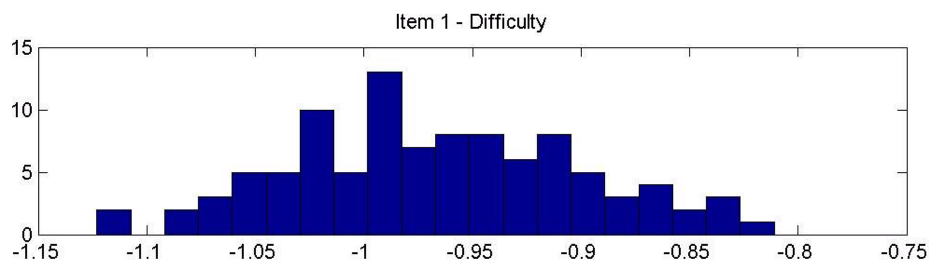

Figure 2: Histogram of the estimated difficulty parameter over many datasets

3. Item bootsrapping

Consider the case with relatively small number of items but relatively large number of cognitive attributes under interest.

Let $I_{n\times p}$ be a matrix of correct (1) and incorrect (0) answers for a set of p items given by n examinees ($I_{ij} = 1$ states that examine i gives a correct answent on item j). Let $Q_{n\times k} = \{q_{ij}\}$ is the corresponding Q-matrix for the attributes $C_1, \ldots, C_k$.

Suppose that we need N items ($N > n$, $N > k$) in order to use equation (5) to calculate the the performance of the attributes. Then some additional (artificial) items should be added to the original set. This means that the item responce for these items should be generated and added to the matrix I. Additionally the matrix Q should be expanded to complete the new set of items.

Let $w_j, i = 1, \ldots, N,\ j = 1, \ldots, k$ are the probabilities that a new artificial item depends on attribute j. These probabilities represent the dependence between items in the test and cognitive attributes. They can be calculated as proportions $w_j = \sum\limits_{i=1}^{p} q_{ij}/p$, or for particular purposes, one can set $w_j = 1/2$.

The first n rows of the new attribute matrix $\tilde{Q}_{N\times k}$ consist of the matrix Q. The other $N-n$ rows can be generated as a results of Bernoulli experiments with probabilities w_j.

The probability P_l; $l = 1, \ldots, n$ for correct answer for the item I_l is calculated as proportion

$$P_l = \frac{\sum\limits_{i=1}^{m} I_{il}}{m}.$$

In other hand, this probability can be calculated using ICC if one knows the IRT parameters of the item.

Then the probability S_j, $j = 1, \ldots, k$ that a student possess a given attribute C_j can be calculated as least squares solution (following Dimitrov, 2007) of

$$P_l = \prod_{j=1}^{k} S_j^{Q_{hj}}.$$

The performance on a new item is genetared as a result of a Bernoulli experiment over the set of attributes $C_1, \ldots, C_k$ with probabilities $S_1, \ldots, S_k$. Then the probability for correct answer on the item I_h, $h > n$ is calculated as

$$R_h = \prod_{j=1}^{k} S_j^{Q_{N_{hj}}}.$$

Then the responce on the new item can be obtained using the algorithm presented in previous section.

4. Conclusions

Proposed methods for bootstrapping the item responces over a set of items or set of attributes, required for correct responce, gives opportunity to artificially increase the number of observations under study.

No additional properties of the estimators should be expected. In order to studied for their accuracy the proposed algorithms should be applied to the set of items with known parameters. Because of the available data it was possible only for the case of small number of observations, described in Section 2.

Data comes from English Language Test with 60 items performed in New Bulgarian University on the 320 second year students. The difficulty parameters $d_1, \ldots, d_{60}$ of the items $I_1, \ldots, I_{60}$ are estimated as usual by a least squares fit to the ogive curve of ICC.

A 100 subsamples S_i, $i = 1, \ldots, 100$ with 30 randomly chosen student responces are subtracted from the original data set. Let us set $M = 500$ and $N = 200$, according the notations in Section 2. The ability scale is set to be $(-3, 3)$.

Then, the difficulty parameter $\hat{d}_{ij}$ for the item I_j over the subset S_i is calculated using the algorithm presented in Section 2. The relative differences for

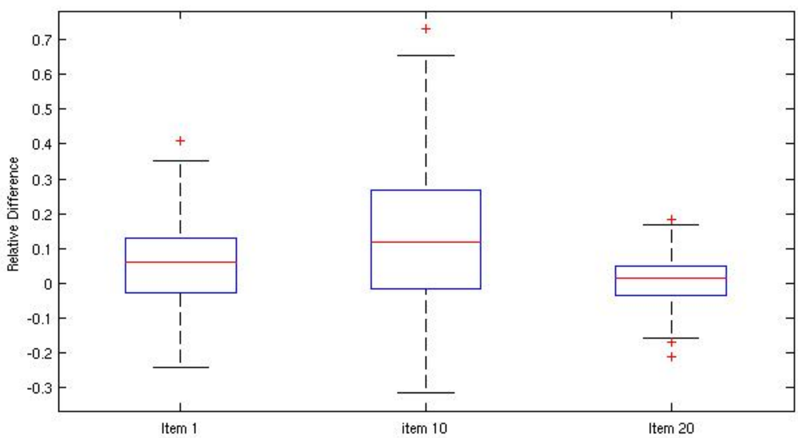

Figure 3: Boxplot of the relative differences

the estimated values $r_{ij} = \dfrac{d_j - \hat{d}_{ij}}{d_j}$ are calculated. The boxplot of these relative differences for the I_1, I_{10} and I_{20} are presented on Fig. 3

Then the total relative difference R_j of the estimates of the difficulty parameter on item I_j is obtained using $R_j = \dfrac{1}{100}\sum_{i=1}^{100} r_{ij}$; $j = 1, \ldots, 60$. The mean value of R_j; $j = 1, \ldots, 60$ is 0.10561 and the variance is 0.1987.

Thus one can assume that the proposed algorithm can be used in everyday practice without general lose of accuracy of the estmated parameters.

Remark: Proposed methods are included in the MATLAB package IRT, available in `http://evanuation.nbu.bg/`.

REFERENCES

[1] M. R. Chernick. Bootstrap Methods, A practitioner's guide. Wiley Series in Probability and Statistics, 1999.

[2] L. Crocker, J. Algila. Introduction to Classical and Modern Test Theory. Warsworth, 1986.

[3] D. Dimitrov. Least Squares Distance Method of Cognitive Validation and Analysis for Binary Items Using Their Item Responce Theory Parameters. Applied Psychological Measurement, 2007.

[4] K. Ghada. The Effects of Sample Size on the Equating of Test Items. Education, 2005.

[5] D. de Gruijter. Small N Does Not Always Justify Rasch Model. *Applied Psychological Measurement* **10** (1986), 187–194.

[6] F. Lord. Applications of Item Responce Theory to Practical Testing Problems. Lawrence Erlbaum Ass. Inc, 1980.

[7] M. Ree, N. Jensen. Item Characteristic Curve Parameters: Effect of Sample Size on Linear Equating. Air Force Human Resources Lab. Report AFHRL-TR-79-70, 1980.

[8] E. Smith, R. Smith. Introduction to Rasch Measurement. JAM Press, Maple Grove, Minesota US, 2004.

[9] M. Stone, F. Yumoto. The Effect of Sample Size For Estimating Rasch/IRT Parameters With Dichotomous Items. *J. Appl. Meas.* **5**, No 1 (2004), 48–61.

Dimitar Atanasov
New Bulgarian University
21, Montevideo Str.
Sofia, Bulgaria
e-mail: datanasov@nbu.bg