Research Report

Latent-Class Item Response Models

Shelby J. Haberman

Research & Development

December 2005
RR-05-28

Latent-Class Item Response Models

Shelby J. Haberman
ETS, Princeton, NJ

December 2005

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Latent-class item response models with small numbers of latent classes are quite competitive in terms of model fit to corresponding item-response models, at least for one- and two-parameter logistic (1PL and 2PL) models. Provided that care is taken in terms of computational procedures and in terms of use of only limited numbers of latent classes, computations are relatively simple in the case of latent classes.

Key words: Log penalty, Rasch model, 2PL model

Acknowledgements

This paper has benefited from conversations with Matthias von Davier and Paul Holland.

Latent-class item response models have been available for some time (Heinen, 1996); however, problems of computation and parameter stability have been major obstacles to their use. These problems appear not to be serious if latent-class item response models are employed with tests of substantial length, if the number of latent classes is small, and if a stabilized Newton-Raphson algorithm is employed for maximum-likelihood estimation (Haberman, 1988). Under these conditions, results are quite competitive to those obtained with item response models in which a normal ability distribution is assumed. These claims are explored in this paper in two cases. The one-parameter logistic (1PL) model is studied in Section 1, and the two-parameter logistic (2PL) model is considered in Section 2. Results are illustrated by use of data from the Praxis™ series of examinations. The criterion of model quality employed is that of estimated expected log penalty per item (Gilula & Haberman, 1994, 1995; Haberman, 2004). Implications of results for psychometric practice are considered in Section 3.

Throughout this report, $n \geq 1$ examinees each take a test with $q \geq 3$ items, a random variable X_{ij} is 1 if item j is answered correctly by examinee i, and X_{ij} is 0 if item j is not answered correctly. Each vector $\mathbf{X}_i$ of responses X_{ij}, $1 \leq j \leq q$, is independent and identically distributed. The set Γ of possible values of $\mathbf{X}_i$ consists of all q-dimensional vectors such that each coordinate is 0 or 1. The distribution of $\mathbf{X}$ is characterized by the array $\mathbf{p}$ of probabilities

$$p(\mathbf{x}) = P(\mathbf{X}_i = \mathbf{x})$$

for $\mathbf{x}$ in Γ, so that $\mathbf{p}$ is in the simplex T of arrays $\mathbf{r}$ with nonnegative elements $r(\mathbf{x})$, $\mathbf{x}$ in Γ, with sum 1. The log likelihood function at $\mathbf{r}$ in T is then

$$\ell(\mathbf{r}) = \sum_{i=1}^{n} \log r(\mathbf{X}_i),$$

and

$$\hat{H}(\mathbf{r}) = -(nq)^{-1}\ell(\mathbf{r})$$

estimates the expected log penalty per item

$$H(\mathbf{r}) = -q^{-1}E(\log r(\mathbf{X}_1))$$

from probability prediction of $\mathbf{X}_1$ by use of $\mathbf{r}$. For a nonempty subset S of T, the maximum log likelihood $\ell(S)$ of $\ell(\mathbf{r})$ for $\mathbf{r}$ in S then leads to the minimum estimated expected log penalty per

item $\hat{H}(S) = -(nq)^{-1}\ell(S)$ of $\hat{H}(\mathbf{r})$ for $\mathbf{r}$ in S. Here $\hat{H}(S)$ is an estimate of the minimum expected log penalty per item $H(S)$ of $H(\mathbf{r})$ for $\mathbf{r}$ in S.

In all models under study, a random ability variable θ_i is associated with each examinee i, and the X_{ij}, $1 \leq j \leq q$, are conditionally independent given θ_i. The pairs $(\theta_i, \mathbf{X}_i)$ are independent and identically distributed, and the distribution function of θ_i is D. For each item j, the conditional probability $P_j(\theta)$ that $X_{ij} = 1$ given $\theta_i = \theta$ is positive and less than 1, so that $Q_j(\theta) = 1 - P_j(\theta)$ is also positive and less than 1. The function P_j is the item characteristic curve, and

$$\lambda_j = \log(P_j/Q_j)$$

is the item logit function (Holland, 1990), so that

$$P_j = \frac{\exp(\lambda_j)}{1 + \exp(\lambda_j)} \tag{1}$$

and

$$Q_j = \frac{1}{1 + \exp(\lambda_j)}. \tag{2}$$

Let $\boldsymbol{\lambda}$ have coordinates λ_j for $1 \leq j \leq q$, and let

$$\mathbf{u}'\mathbf{v} = \sum_{j=1}^{q} u_j v_j$$

for q-dimensional vectors u and v with respective coordinates u_j and v_j for $1 \leq j \leq q$. For

$$V = \prod_{j=1}^{q} Q_j = \prod_{j=1}^{q} \frac{1}{1 + \exp(\lambda_j)}, \tag{3}$$

a variation on the Dutch identity yields

$$p(\mathbf{x}) = \int V \exp(\mathbf{X}_i'\boldsymbol{\lambda}) dD \tag{4}$$

(Holland, 1990).

For the data under study, $q = 45$ and $n = 8,686$. It is helpful in the analysis that both the number of items q and the number of examinees n are both relatively large. The relatively large sample size contributes to stability of estimates. The relatively large number of items appears to permit more latent categories to be used (Haberman, 2005). In the latent-class item response models considered in this report, a fixed finite set of possible values τ_k, $1 \leq k \leq K$, of θ_i is given for some integer $K \geq 2$. The probability that $\theta_i = \tau_k$ is

$$\frac{\exp(\nu_k)}{\sum_{k'=1}^{K} \exp(\nu_{k'})}$$

for $1 \leq k \leq K$, where the constraint $\sum_{k=1}^{K} \nu_k = 0$ is used to identify parameters. Thus

$$p(\mathbf{x}) = \frac{\sum_{k=1}^{K} V(\tau_k) \exp(\mathbf{X}_i' \boldsymbol{\lambda}(\tau_k) + \nu_k)}{\sum_{k=1}^{K} \exp(\nu_k)} \tag{5}$$

for all $\mathbf{x}$ in Γ.

The stabilized Newton-Raphson algorithm (Haberman, 1988) can be employed with little difficulty in cases in which, conditional on θ, λ_j is a linear function of a vector of unknown parameters independent of the ν_k. The algorithm employed here is slightly modified from the previously reported version due to the presence of inequality restraints and due to the ease with which the Hessian matrix can be computed relative to the ease with which the information matrix can be computed (Haberman, 1988). The changes render the algorithm rather similar to a variant on the Newton-Raphson algorithm used earlier for log-linear models (Haberman, 1974, ch. 3). Use of the stabilized Newton-Raphson algorithm provides estimated asymptotic standard deviations of parameter estimates as a direct result of the computations, so that it permits a quite straightforward analysis of data relative to that provided by the EM algorithm. In addition, very large estimated asymptotic standard deviations provide evidence of estimation problems.

1. The 1PL Model

In the case of the 1PL model,

$$\lambda_j = a\theta - \gamma_j \tag{6}$$

for unknown parameters $a > 0$ and γ_j. The set S_{1K} for this model consists of members $\mathbf{p}$ in T such that (5) holds for all $\mathbf{x}$ in Γ, (6) holds for some $a > 0$ and γ_j for $1 \leq j \leq q$, and (3) holds. The common item discrimination is a and the item difficulty is $\beta_j = \gamma_j / a$. For this case, evenly spaced values of τ_k were considered for each K tried. To simplify comparisons, the τ_k were arranged so that θ_i would have mean 0 and variance 1 if θ_i were uniformly distributed on the τ_k. Values of K from 2 to 5 were examined. Results are summarized in Table 1 for the Praxis example. It should be noted that use of a less restricted conditional Rasch model in which no assumptions are made concerning D yields an estimated expected log penalty per item of 0.59611 (Haberman, 2004). This estimate for the case of D unrestricted implies that no latent-class 1PL model can yield an estimated expected penalty less than 0.59611, so that it is clearly not possible to improve much upon the latent-class 1PL model with five classes.

Table 1.
Estimated Expected Log Penalties for Observations for IRT Models

Model type	Number of classes	Estimated expected log penalty per item
Normal 1PL		0.59639
Latent-class 1PL	2	0.60075
Latent-class 1PL	3	0.59716
Latent-class 1PL	4	0.59641
Latent-class 1PL	5	0.59621
Normal 2PL		0.59157
Latent-class 2PL	2	0.59708
Latent-class 2PL	3	0.59269
Latent-class 2PL	4	0.59164
Latent-class 2PL	5	0.59124

Table 2.
Maximum Differences Between Empirical and Estimated Distribution Functions of the Score Sum for IRT Models

Model type	Number of classes	Maximum difference
Normal 1PL		0.0302
Latent-class 1PL	2	0.0635
Latent-class 1PL	3	0.0313
Latent-class 1PL	4	0.0154
Latent-class 1PL	5	0.0071
Normal 2PL		0.0410
Latent-class 2PL	2	0.0575
Latent-class 2PL	3	0.0371
Latent-class 2PL	4	0.0195
Latent-class 2PL	5	0.0105

Note that four or five latent classes yields a model quite comparable to the normal 1PL model in terms of estimated expected log penalty per observation. More than five latent classes leads to poor identification of parameters and only limited model improvement, a result not unexpected given known problems with latent-class models (Haberman, 2005). Bias in estimation of expected log penalty per observation is a rather small problem given the large sample size, so corrections for bias are not considered in this report. Corrections are available for smaller sample sizes (Gilula & Haberman, 2001).

It is of some interest to look at the estimated and observed marginal distributions of the score sum $X_{i+} = \sum_{j=1}^{q} X_{ij}$ to consider how well the various models approximate this distribution. Results are summarized in Table 2. The measure employed is the maximum difference $D(S)$ between the sample and estimated distribution functions of X_{i+} for model S. By this criterion, the models with four or five latent classes are somewhat more effective than is the customary normal 1PL model. Nonetheless none of the discrepancies of distribution functions is especially large. To understand observed differences in cumulative distribution functions, observe that two normal distributions, each with a standard deviation of 1, have distribution functions that differ by as much as 0.08 if the means differ by 0.2, while the distribution functions differ by as much as 0.03 if the means differ by 0.04.

2. The 2PL Model

In the case of the 2PL model,

$$\lambda_j = a_j\theta - \gamma_j \tag{7}$$

for unknown parameters $a_j > 0$ and γ_j. The set S_{2K} for this model consists of members $\mathbf{p}$ in T such that (5) holds for all $\mathbf{x}$ in Γ, (7) holds for some $a_j > 0$ and γ_j for $1 \leq j \leq q$, and (3) holds. The item discrimination is a_j and the item difficulty is $\beta_j = \gamma_j/a_j$. The same choice of τ_k was made as it was for the 1PL case for K from 2 to 5. Again the Praxis data were used to obtain the results summarized in Table 1. The situation is quite similar to that observed for the 1PL case. Four to five latent classes yield results quite comparable to those for the normal 2PL model. Once again, more latent classes do not help appreciably. Note that even a 2PL model with only three latent classes is more successful than any 1PL model in terms of estimated expected log penalty per item. Note that the comparison of empirical and estimated distribution functions of X_{i+} in Table 2 favors latent-class models with four or five latent classes over the normal 2PL model.

3. Conclusions

The results in this report can be interpreted in two directions. On the one hand, latent-class 1PL and 2PL models do not appear to offer much improvement over normal 1PL and 2PL models, at least for the example under study. On the other hand, it is possible to obtain results quite competitive with a normal 1PL or 2PL model with remarkably few latent classes. This result may reasonably be regarded as disturbing. A standard normal distribution is not well-approximated by

a discrete distribution with four or five values. Nonetheless, the observations are described about equally well by use of a normal ability distribution or by use of an ability distribution confined to four or five equally spaced points. It follows that little empirical evidence exists concerning the nature of the ability distribution even if one believes that the 2PL model or 1PL model really holds. This lack of evidence concerning the nature of the ability distribution in turn implies that estimation of an ability parameter for an individual examinee is rather problematic. The posterior distribution of the ability variable θ_i given the observed response vector $\mathbf{X}_i$ is quite different if the ability distribution is normal than if the ability distribution is confined to five equally spaced points. This problem is not an immediate issue within the Praxis program, for scoring does not employ item response theory, but no obvious reason exists for the issues addressed here not to arise in other examinations as well.

References

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656.

Gilula, Z., & Haberman, S. J. (1995). Prediction functions for categorical panel data. *The Annals of Statistics*, *23*, 1130–1142.

Gilula, Z., & Haberman, S. J. (2001). Analysis of categorical response profiles by informative summaries. *Sociological Methodology*, *31*, 129–187.

Haberman, S. J. (1974). *The analysis of frequency data*. Chicago: University of Chicago Press.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology*, *18*, 193–211.

Haberman, S. J. (2004). *Maximum likelihood for the Rasch model for binary responses* (ETS RR-04-20). Princeton, NJ.

Haberman, S. J. (2005). *Identifiability of parameters in item response models with unconstrained ability distributions* (ETS RR-05-24). Princeton, NJ.

Heinen, T. (1996). *Latent class and discrete latent trait models*. Thousand Oaks, CA: Sage.

Holland, P. W. (1990). The Dutch identity: a new tool for the study of item response models. *Psychometrika*, *55*, 5–18.