Research Report

Use of Generalized Residuals to Examine Goodness of Fit of Item Response Models

Shelby J. Haberman

April 2009

ETS RR-09-15

Use of Generalized Residuals to Examine Goodness of Fit of Item Response Models

Shelby J. Haberman

ETS, Princeton, New Jersey

April 2009

Abstract

Generalized residuals are a tool employed in the analysis of contingency tables to examine goodness of fit. They may be applied to item response models with little complication. Their use is illustrated with testing data from operational programs. Models considered include the Rasch model and the two-parameter logistic model.

Key words: Marginal distributions, normal approximations, contingency tables, conditional estimation, marginal estimation, Rasch model, two-parameter logistic model

Acknowledgments

The author thanks Neil Dorans, Dan Eignor, Frank Rijmen, Sandip Sinharay, and Matthias von Davier for their helpful comments.

Generalized residuals have been applied in the analysis of contingency tables to examine the goodness of fit of log-linear (Haberman, 1976, 1978, 1979) and latent-class models (Haberman, 1979, chap. 10). The essential feature of these residuals is that a linear combination O of observed frequencies is compared to its estimated expected value $\hat{E}$ under a proposed model. Under the assumption that the model holds, the standard deviation of the difference $O - \hat{E}$ is then estimated by a consistent estimator s. If the model holds, and if the sample size is large, then it is typically true that the generalized residual $t = (O - \hat{E})/s$ has an approximate standard normal distribution, so that a large value of $|t|$ indicates a failure of the model under study. Variations on residuals of this kind are quite old in special cases (Armitage, 1955; Cochran, 1954, 1955; Mantel & Haenszel, 1959; Yates, 1948).

Within item-response analysis, applications of generalized residuals have been quite rare. Exceptions are found in the case of the Rasch model (Glas & Verhelst, 1989, 1995; Haberman, 2004). Most residuals that have been employed in item-response analysis are standardized residuals of the form $t_0 = (O - \hat{E})/s_0$, where s_0 is an estimate of the standard deviation of the observed linear combination O (Hambleton, Swaminathan, & Rogers, 1991). Standardized residuals are generally easier to compute than are generalized residuals; however, even if the model holds and samples are large, the approximate distribution of t_0 is a normal distribution with mean 0 but variance less than 1. As a consequence, standardized residuals, if treated as standard normal random variables, are likely to suggest model agreement, when no such agreement exists. An additional complication, in practice, has been use of residuals in which the linear combination O is not an observed quantity at all. For example, a common use of a standardized residual to examine the fit of items to a model employs an O that is a joint frequency in which an observed variable has a specified value and an unobserved variable is in a specified range.

To eliminate the deficiencies of residual analysis in item-response analysis requires attention to proper estimation of the standard deviation of the difference $O - \hat{E}$ and proper attention to quantities that are actually observed. This effort is considered in this report for a common testing situation in which n examinees each respond to q right-scored items. For examinee i, $1 \leq i \leq n$, and item j, $1 \leq j \leq q$, the scored response is denoted by X_{ij}, where X_{ij} is 1 for a correct answer and 0 otherwise. The vector of responses X_{ij}, $1 \le j \le q$, for examinee i is then $\mathbf{X}_i$. It is assumed that the vectors $\mathbf{X}_i$ are independent and identically distributed. For each $\mathbf{x}$ in the set Γ of possible response vectors, the probability that $\mathbf{X}_i = \mathbf{x}$ is denoted by $p(\mathbf{x}) > 0$. If $n(\mathbf{x})$ denotes the number of examinees i, $1 \le i \le n$, with $\mathbf{X}_i = \mathbf{x}$, then a simple estimate of $p(\mathbf{x})$ is $\bar{p}(\mathbf{x}) = n(\mathbf{x})/n$. For some real function d on Γ, the linear combination under study is

$$O = \sum_{\mathbf{x}\in\Gamma} d(\mathbf{x})\bar{p}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} d(\mathbf{X}_i). \tag{1}$$

The expected value of O is

$$E = \sum_{\mathbf{x}\in\Gamma} d(\mathbf{x})p(\mathbf{x}) = E(d(\mathbf{X}_i)). \tag{2}$$

For the model under study, $p(\mathbf{x})$ has an estimate $\hat{p}(\mathbf{x})$, so that

$$\hat{E} = \sum_{\mathbf{x}\in\Gamma} d(\mathbf{x})\hat{p}(\mathbf{x}). \tag{3}$$

For all cases under study, item-response models are considered with the local independence assumption that

$$p(\mathbf{x}) = E(p(\mathbf{x}|\theta)), \tag{4}$$

where θ is a random variable with distribution F,

$$p(\mathbf{x}|\theta) = \prod_{j=1}^{q} p_j(x_j|\theta), \tag{5}$$

and

$$p_j(1|\theta) + p_j(0|\theta) = 1. \tag{6}$$

The three-parameter logistic (3PL) model is considered, for which

$$p_j(1|\theta) = c_j + (1 - c_j)\frac{\exp(a_j\theta - \beta_j)}{1 + \exp(a_j\theta - \beta_j)}; \tag{7}$$

a_j, β_j, and c_j are real; and $0 \le c_j < 1$ (Hambleton et al., 1991, chap. 1). For item j, c_j is the guessing parameter. In the customary case with a_j positive, a_j is the item discrimination and β_j/a_j is the item difficulty. In a one-parameter logistic (1PL) model (Rasch, 1960), a_j is assumed to be constant and c_j is assumed to be 0. In a two-parameter logistic (2PL) model, c_j is assumed to be 0. In a 3PL model, no restrictions are imposed on a_j, β_j, or c_j. Results are illustrated in all cases by the use of an administration of a test for prospective teachers that was previously considered in Haberman (2005).

In the conditional estimation case of section 2, the Rasch version of the 1PL model is assumed, so that no restrictions are imposed on the distribution F. In this case, generalized residuals derived from log-linear models may be applied (Haberman, 1976, 1978, 1979, 2004), for the $\mathbf{X}_i$ satisfy the log-linear model

$$\log p(\mathbf{x}) = \gamma_{S(\mathbf{x})} - \boldsymbol{\beta}'\mathbf{x}, \tag{8}$$

where $S(\mathbf{x}) = \sum_{j=1}^{q} x_j$ is the sum of the coordinates of $\mathbf{x}$, $\boldsymbol{\beta}$ is the q-dimensional vector of β_j, $1 \le j \le q$, and

$$\boldsymbol{\beta}'\mathbf{x} = \sum_{j=1}^{q} \beta_j x_j$$

(Tjur, 1982). In this case, and in generalizations to the partial credit model, some special instances of generalized residuals have appeared in the literature (Glas & Verhelst, 1989; Haberman, 2004). In addition, the square of a generalized residual for the Rasch model yields a generalized Pearson test (Glas & Verhelst, 1995).

In section 3, a 2PL model is considered, in which the distribution F is assumed to be a standard normal distribution (Bock & Aitkin, 1981). Here procedures in Haberman (1979) for log-linear models for indirectly observed contingency tables cannot be applied directly because θ_i has a normal distribution. In addition, further complications arise in practice because 2^q, the number of elements in Γ, can be a very large number.

Implications of results are considered in section 4.

2 Conditional Estimation

Conditional estimation is commonly applied to the Rasch case with ability distribution F unknown (Andersen, 1970, 1972). In this case, the model, together with the probabilities $p(\mathbf{x})$, $\mathbf{x}$ in Γ, do not determine F and $\boldsymbol{\beta}$; however, it is quite adequate for the analysis to require that, for some fixed q-dimensional vector $\mathbf{c}$ with coordinate sum $\sum_{j=1}^{q} c_j \neq 0$, $\mathbf{c}'\boldsymbol{\beta} = 0$. For $0 \leq k \leq q$, let $\Gamma(k)$ denote the set of response vectors $\mathbf{x}$ in Γ such that $S(\mathbf{x}) = k$. Under the model, if $\mathbf{x}$ is an element of $\Gamma(k)$ for an integer k such that $0 \leq k \leq q$, and if

$$T_k(\boldsymbol{\beta}) = \sum_{\mathbf{x}\in\Gamma(k)} \exp(-\boldsymbol{\beta}'\mathbf{x}), \tag{9}$$

then

$$p(\mathbf{x};\boldsymbol{\beta}) = [T_k(\boldsymbol{\beta})]^{-1} \exp(-\boldsymbol{\beta}'\mathbf{x}) \tag{10}$$

is the conditional probability that $\mathbf{X}_i = \mathbf{x}$, given that $S_i = S(\mathbf{X}_i) = k$. Let

$$\mathbf{T}_{k1}(\boldsymbol{\beta}) = \sum_{\mathbf{x}\in\Gamma(k)} \mathbf{x} \exp(-\boldsymbol{\beta}'\mathbf{x}). \tag{11}$$

The conditional expectation $\mathbf{m}_k(\boldsymbol{\beta})$ of $\mathbf{X}_i$, given $S_i = k$, is then

$$\mathbf{m}_k(\boldsymbol{\beta}) = [T_k(\boldsymbol{\beta})]^{-1}\mathbf{T}_{k1}(\boldsymbol{\beta}). \tag{12}$$

Let

$$\mathbf{T}_{k2}(\boldsymbol{\beta}) = \sum_{\mathbf{x}\in\Gamma(k)} [\mathbf{x} - \mathbf{m}_k(\boldsymbol{\beta})][\mathbf{x} - \mathbf{m}_k(\boldsymbol{\beta})]' \exp(-\boldsymbol{\beta}'\mathbf{x}). \tag{13}$$

The conditional covariance matrix $\mathbf{V}_k(\boldsymbol{\beta})$ of $\mathbf{X}_i$, given $S_i = k$, is

$$\mathbf{V}_k(\boldsymbol{\beta}) = [T_k(\boldsymbol{\beta})]^{-1}\mathbf{T}_{k2}(\boldsymbol{\beta}). \tag{14}$$

Let $n_S(k)$ denote the number of examinees i for whom $S_i = k$, let $p_S(k)$ be the probability that $S_i = k$, and let $\hat{p}_S(k)$ denote the fraction $n_S(k)/n$ of examinees i for whom $S_i = k$. Then the conditional maximum-likelihood estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, if it exists, satisfies the equations $\mathbf{c}'\hat{\boldsymbol{\beta}} = 0$, $\hat{\mathbf{m}}_k = \mathbf{m}_k(\hat{\boldsymbol{\beta}})$ for $0 \leq k \leq q$, and

$$\sum_{k=0}^{q} \hat{p}_S(k)\hat{\mathbf{m}}_k = n^{-1}\sum_{i=1}^{n} \mathbf{X}_i. \tag{15}$$

Note that coordinate j of $n^{-1}\sum_{i=1}^{n} \mathbf{X}_i$ is the fraction f_{j+} of examinees i with $X_{ij} = 1$. These equations determine $\hat{\boldsymbol{\beta}}$ uniquely if $n_S(0) + n_S(q) < n$. As the sample size n increases, the probability approaches 1 that $\hat{\boldsymbol{\beta}}$ is uniquely defined, and $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a multivariate normal random vector with zero mean and with covariance matrix

$$\mathbf{C} = (\mathbf{V} + \mathbf{c}\mathbf{c}')^{-1}\mathbf{V}(\mathbf{V} + \mathbf{c}\mathbf{c}')^{-1}, \tag{16}$$

where

$$\mathbf{V} = \sum_{k=0}^{q} p_S(k)\mathbf{V}_k(\boldsymbol{\beta}). \tag{17}$$

Let $\hat{\mathbf{V}}_k = \mathbf{V}_k(\hat{\boldsymbol{\beta}})$. One may estimate $\mathbf{V}$ by

$$\hat{\mathbf{V}} = \sum_{k=0}^{q} \bar{p}_S(k)\hat{\mathbf{V}}_k, \tag{18}$$

so that $\mathbf{C}$ may be estimated by

$$\hat{\mathbf{C}} = (\hat{\mathbf{V}} + \mathbf{c}\mathbf{c}')^{-1}\hat{\mathbf{V}}(\hat{\mathbf{V}} + \mathbf{c}\mathbf{c}')^{-1}. \tag{19}$$

The corresponding estimate of $p(\mathbf{x})$ is then

$$\hat{p}(\mathbf{x}) = \sum_{k=0}^{q} \hat{p}_S(k)p(\mathbf{x}; \hat{\boldsymbol{\beta}}).$$

For a real function d on Γ, the observed mean O and its estimated mean $\hat{E}$ may now be defined as in (1) and (3). For useful results, assume that $d(\mathbf{x})$ is not 0 for all $\mathbf{x}$ in Γ; and no γ_k, $0 \le k \le q$; and q-dimensional vectors $\boldsymbol{\alpha}$ exist such that

$$d(\mathbf{x}) = \gamma_k - \boldsymbol{\alpha}'\mathbf{x} \tag{20}$$

for all $\mathbf{x}$ in $\Gamma(k)$ and for all integers k, $0 \le k \le q$. This assumption is needed to avoid trivial cases in which O and $\hat{E}$ are the same whenever $\hat{\boldsymbol{\beta}}$ exists.

A normal approximation for $O - \hat{E}$ is typically available when the model holds (Haberman, 1978, 2004). Let D_k be the mean

$$D_k = \sum_{\mathbf{x}\in\Gamma(k)} d(\mathbf{x})p(\mathbf{x}; \boldsymbol{\beta}) \tag{21}$$

of $d(\mathbf{X}_i)$, given $S_i = k$. Let $\boldsymbol{\eta}$ be the expected conditional covariance

$$\boldsymbol{\eta} = \sum_{k=0}^{q} p_S(k) \sum_{\mathbf{x}\in\Gamma(k)} [d(\mathbf{x}) - D_k][\mathbf{x} - \mathbf{m}_k(\boldsymbol{\beta})]p(\mathbf{x}; \boldsymbol{\beta}) \tag{22}$$

of $d(\mathbf{X}_i)$ and $\mathbf{X}_i$, given S_i. Let

$$\boldsymbol{\alpha} = (\mathbf{V} + \mathbf{c}\mathbf{c}')^{-1}\boldsymbol{\eta}, \tag{23}$$

so that the expected conditional variance of $d(\mathbf{X}_i) - \boldsymbol{\gamma}'\mathbf{X}_i$, given S_i, is minimized subject to $\mathbf{c}'\boldsymbol{\gamma} = 0$ if $\boldsymbol{\gamma} = \boldsymbol{\alpha}$. Let the expected conditional variance of $d(\mathbf{X}_i) - \boldsymbol{\alpha}'\mathbf{X}_i$, given S_i, be denoted by

$$\tau^2 = \sum_{k=0}^{q} p_S(k) \sum_{\mathbf{x}\in\Gamma} \{d(\mathbf{x}) - D_k - \boldsymbol{\alpha}'[\mathbf{x} - m_k(\boldsymbol{\beta})]\}^2 p(\mathbf{x};\boldsymbol{\beta}). \tag{24}$$

Then $z = n^{1/2}(O - \hat{E})/\tau$ has an approximate standard normal distribution. The approximation can be made increasingly accurate by requiring that each ratio $d(\mathbf{x})/(n^{1/2}\tau)$ be small and that each element of $(n\mathbf{C})^{-1}$ be small.

To estimate τ^2, let

$$\hat{D}_k = \sum_{\mathbf{x}\in\Gamma(k)} d(\mathbf{x})\hat{p}(\mathbf{x}), \tag{25}$$

let

$$\hat{\boldsymbol{\eta}} = \sum_{k=0}^{q} \hat{p}_S(k) \sum_{\mathbf{x}\in\Gamma(k)} [d(\mathbf{x}) - \hat{D}_k](\mathbf{x} - \hat{\mathbf{m}}_k)\hat{p}(\mathbf{x}), \tag{26}$$

let

$$\hat{\boldsymbol{\alpha}} = (\hat{\mathbf{V}} + \mathbf{c}\mathbf{c}')^{-1}\hat{\boldsymbol{\eta}}, \tag{27}$$

and let

$$\hat{\tau}^2 = \sum_{k=0}^{q} \hat{p}_S(k) \sum_{\mathbf{x}\in\Gamma} [d(\mathbf{x}) - \hat{D}_k - \hat{\boldsymbol{\alpha}}'(\mathbf{x} - \hat{\mathbf{m}}_k)]^2 \hat{p}(\mathbf{x}). \tag{28}$$

An equivalent but different formula has been previously provided by the author (Haberman, 2004). If $s = \hat{\tau}/n^{1/2}$, then it is typically true that $t = (O - \hat{E})/s$ has an approximate standard normal distribution if the Rasch model holds, so that large values of $|t|$ suggest incompatibility of the model with the data. The requirements for the normal approximation for t are the same as for z.

Formulas are simplest if $d(\mathbf{x}) = g(S(x))\mathbf{u}'\mathbf{x}$ for some real function g on the integers 0 to q and some q-dimensional vector $\mathbf{u}$ with coordinates that are not constant. For useful results, it must be the case that $g(k)$ is not constant for $1 \le k \le q - 1$. Let f_{jk} be the fraction of examinees i with $S_i = k$ and $X_{ij} = 1$, let $f_{j\cdot k} = f_{jk}/\hat{p}_S(k)$ if $\hat{p}_S(k) > 0$, and let $f_{j\cdot k}$ be f_{j+} if $\hat{p}_S(k) = 0$. Let $\mathbf{f}_k$ be the q-dimensional vector with coordinates $f_{j\cdot k}$ for $1 \le j \le q$. Let $\hat{\mathbf{m}}_{kj}$ be $\mathbf{m}_k(\hat{\boldsymbol{\beta}})$. Then

$$O = \sum_{k=0}^{q} g(k)\hat{p}_S(k)\mathbf{u}'\mathbf{f}_k, \tag{29}$$

and

$$\hat{E} = \sum_{k=0}^{q} g(k)\hat{p}_S(k)\mathbf{u}'\hat{\mathbf{m}}_k. \tag{30}$$

Given (12), if

$$\bar{g} = n^{-1}\sum_{i=1}^{n} g(S_i) = \sum_{k=0}^{q} g(k)\hat{p}_S(k), \tag{31}$$

then

$$O - \hat{E} = \sum_{k=0}^{q} [g(k) - \bar{g}]\hat{p}_S(k)\mathbf{u}'(\mathbf{f}_k - \hat{m}_k). \tag{32}$$

In this case,

$$\hat{\boldsymbol{\eta}} = \sum_{k=0}^{q} \hat{p}_S(k)g(k)\hat{\mathbf{V}}_k\mathbf{u}. \tag{33}$$

Given $\mathbf{u}$, computation of $\hat{\boldsymbol{\alpha}} = \hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\eta}}$ requires only $g(k)$, $\hat{p}_S(k)$, and $\hat{\mathbf{V}}_k$ for $0 \le k \le q$. It also follows that if

$$\hat{\mathbf{h}}_k = g(k)\mathbf{u} - \hat{\boldsymbol{\alpha}},\ 1 \le k \le q, \tag{34}$$

then

$$\hat{\tau}^2 = \sum_{k=0}^{q} \hat{p}_S(k)\hat{\mathbf{h}}_k'\hat{\mathbf{V}}_k\hat{\mathbf{h}}_k. \tag{35}$$

In many applications, for some item j, $\mathbf{u}$ has coordinate j equal to 1 and all other coordinates equal to 0 (Haberman, 2004). In this case,

$$O = \sum_{k=0}^{q} g(k)\hat{p}_S(k)f_{j\cdot k} \tag{36}$$

is the sample mean of the products $g(S_i)X_{ij}$, $1 \le i \le n$. Similarly,

$$\hat{E} = \sum_{k=0}^{q} g(k)\hat{p}_S(k)\hat{m}_{kj} \tag{37}$$

is the estimated expectation $E(g(S_i)X_{ij})$ of $g(S_i)X_{ij}$. In the remaining computations of the generalized residual, it is helpful to note that $\hat{\mathbf{V}}_j\mathbf{u}$ is then column j of $\hat{\mathbf{V}}_k$.

As an aid to the interpretation of results, it is helpful to consider regression and correlation coefficients. Let

$$\hat{s}_g^2 = n^{-1}\sum_{i=1}^{n}[g(S_i)-\bar{g}]^2 = \sum_{k=0}^{q}[g(k)-\bar{g}]^2\hat{p}_S(k) \tag{38}$$

be the sample variance of $g(S_i)$, $1 \leq i \leq n$. Without use of an item-response model, the linear regression of the item score X_{ij} on the transformed total score $g(S_i)$ has a slope that may be estimated by

$$b = (O - \bar{g}f_{j+})/\hat{s}_g^2. \tag{39}$$

If the Rasch model applies, then the same regression slope can be estimated by

$$\hat{b} = (\hat{E} - \bar{g}f_{j+})/\hat{s}_g^2. \tag{40}$$

Conditional on the observed total scores S_i, $1 \leq i \leq n$, the difference $b - \hat{b} = (O - \hat{E})/\hat{s}_g^2$ in the estimated slopes has standard deviation $n^{-1/2}\hat{\tau}/\hat{s}_g^2$, so that $t = n^{1/2}(O - \hat{E})/\hat{\tau}$ also provides a test of whether the regression of the item score X_{ij} on the transformed total score $g(S_i)$ is appropriately predicted by use of the Rasch model.

In a similar fashion, $O - \hat{E}$ may be used to compare the observed and fitted point-biserial correlations of the item score X_{ij} and the transformed total score $g(S_i)$. The ordinary sample statistic is $r = (O - \bar{g}f_{j+})/[\hat{s}_g^2 f_{j+}(1 - f_{j+})]^{1/2}$ and the estimate based on the Rasch model is $\hat{r} = (\hat{E} - \bar{g}f_{j+})/[\hat{s}_S^2 f_{j+}(1 - f_{j+})]^{1/2}$, so that $r - \hat{r} = (O - \hat{E})/[\hat{s}_g^2 f_{j+}(1 - f_{j+})]^{1/2}$. If the Rasch model holds, then $[nf_{j+}(1 - f_{j+})]^{1/2}(r - \hat{r}) = t$ has an approximate standard normal distribution.

To illustrate results, the case of $g(k) = (k - q/2)/q$, $0 \leq k \leq q$, is considered. With this selection for g, $g(S_i)$ is a linear transformation of S_i with a range from -1 to 1. Data are used from an assessment for prospective teachers previously analyzed in a study of model agreement (Haberman, 2005). It should be noted that the assessment does not employ item-response theory in practice, so that results here do not affect current statistical analysis of this assessment. In this example, the number of items q is 45, and the sample size n is 8,686. A summary of results appears in Table 1. Table 1 emphasizes the regression slopes and the point-biserial correlations because these quantities are easier to understand than the original O and $\hat{E}$ statistics. As is clear from Table 1, the Rasch model is quite unsuccessful at fitting the observed relationship between total score and item response. Items are frequently present with sample point-biserial correlations that are either much larger or much smaller than predicted by the Rasch model. This result is consistent with previously reported results for data from the SAT® program (Haberman, 2004).

3 The Two-Parameter Logistic Model With Polytomous Ability Distribution

Analysis of the 2PL model with a standard normal ability distribution for θ_i is relatively straightforward; however, some changes from procedures in Haberman (1979) are needed due to the continuity of θ_i and due to the very large number of elements typically found in Γ. To describe procedures, let $\hat{a}_j$ denote the maximum-likelihood estimate of the item discrimination a_j and let $\hat{\beta}_j$ denote the maximum-likelihood estimate of the item intercept β_j. Let $\boldsymbol{\gamma}$ be the $2q$-dimensional vector with coordinates $\gamma_j = a_j$ and $\gamma_{q+j} = \beta_j$ for $1 \le j \le q$. Let $\hat{\boldsymbol{\gamma}}$ be the maximum-likelihood estimate of $\boldsymbol{\gamma}$, so that $\hat{\gamma}_j = \hat{a}_j$ and $\hat{\gamma}_{q+j} = \hat{\beta}_j$ for $1 \le j \le q$.

Then

$$p_j(h|\theta;\boldsymbol{\gamma}) = \frac{\exp[h(\gamma_j\theta - \gamma_{q+j})]}{1+\exp(\gamma_j\theta - \gamma_{q+j})} \tag{41}$$

is the conditional probability $p_j(h|\theta)$ that $X_{ij} = h$, $h = 0$ or 1, given that $\theta_i = \theta$. The conditional probability that $\mathbf{X}_i = \mathbf{x}$, given that $\theta_i = \theta$, is then

$$p(\mathbf{x}|\theta;\boldsymbol{\gamma}) = \prod_{j=1}^{q} p_j(x_j|\theta;\boldsymbol{\gamma}). \tag{42}$$

Let ϕ be the density function of the standard normal distribution. Then

$$p(\mathbf{x};\boldsymbol{\gamma}) = \int p(\mathbf{x}|\theta;\boldsymbol{\gamma})\phi(\theta)d\theta \tag{43}$$

is the probability that $\mathbf{X}_i = \mathbf{x}$.

For h equal to 0 or 1, $\hat{p}_j(h|\theta) = p_j(h|\theta;\hat{\boldsymbol{\gamma}})$ is the maximum-likelihood estimate of the conditional probability $p_j(h|\theta)$ that $X_{ij} = h$, given that $\theta_i = \theta$. For $\mathbf{x}$ in Γ,

Table 1

Generalized Residuals for the Rasch Model

Item	Observed slope	Fitted slope	Observed point-biserial	Fitted point-biserial	Generalized residual
1	0.616	1.067	0.201	0.348	−15.802
2	0.542	1.132	0.170	0.355	−20.498
3	1.285	1.141	0.398	0.353	5.007
4	1.276	1.043	0.410	0.335	8.278
5	1.286	1.134	0.402	0.355	5.254
6	0.853	0.666	0.341	0.266	7.479
7	0.519	1.080	0.168	0.349	−19.606
8	0.633	0.801	0.244	0.309	−6.244
9	1.277	1.123	0.403	0.354	5.357
10	1.091	0.893	0.377	0.309	7.267
11	0.829	0.927	0.281	0.315	−3.570
12	1.302	1.086	0.420	0.350	7.537
13	1.093	1.050	0.350	0.336	1.513
14	0.878	0.890	0.319	0.323	−0.419
15	0.632	0.847	0.224	0.301	−8.044
16	0.604	0.372	0.324	0.199	11.476
17	0.721	1.130	0.224	0.351	−14.251
18	0.935	1.138	0.292	0.355	−7.041
19	1.519	1.062	0.484	0.338	16.131
20	0.812	1.086	0.256	0.342	−9.637
21	0.933	1.063	0.297	0.338	−4.575
22	1.125	1.069	0.366	0.348	1.960
23	0.729	0.572	0.315	0.247	6.638
24	0.779	0.983	0.257	0.324	−7.332
25	1.035	0.973	0.343	0.323	2.224
26	1.246	1.022	0.417	0.342	7.865
27	0.848	1.116	0.264	0.348	−9.379
28	1.441	1.114	0.449	0.347	11.437
29	0.577	1.057	0.189	0.347	−16.819
30	0.858	1.141	0.266	0.353	−9.838
31	0.985	0.782	0.364	0.288	7.775
32	0.746	0.731	0.285	0.279	0.603
33	1.014	0.863	0.356	0.303	5.604
34	1.309	1.134	0.406	0.351	6.108
35	0.972	0.945	0.341	0.331	0.995
36	1.461	1.144	0.453	0.355	11.005

(Table continues)

Table 1 (continued)

Item	Observed slope	Fitted slope	Observed point-biserial	Fitted point-biserial	Generalized residual
37	0.995	1.112	0.316	0.353	−4.091
38	1.117	0.796	0.409	0.291	12.180
39	1.493	1.138	0.466	0.355	12.327
40	1.284	1.073	0.407	0.340	7.434
41	0.786	0.973	0.271	0.335	−6.627
42	1.123	1.134	0.348	0.352	−0.385
43	0.979	1.120	0.309	0.354	−4.883
44	1.221	1.131	0.383	0.355	3.149
45	1.238	1.144	0.384	0.355	3.241

$\hat{p}(\mathbf{x}|\theta) = p(\mathbf{x}|\theta;\hat{\boldsymbol{\gamma}})$ is the maximum-likelihood estimate of the conditional probability $p(\mathbf{x}|\theta)$ that $\mathbf{X}_i = \mathbf{x}$, given $\theta_i = \theta$, and $\hat{p}(\mathbf{x}) = p(\mathbf{x};\hat{\boldsymbol{\gamma}})$ is the maximum-likelihood estimate of the probability $p(\mathbf{x})$ that $\mathbf{X}_i = \mathbf{x}$. Given these definitions, define O by (1) and $\hat{E}$ by (3).

Estimation of the standard deviation of $O - \hat{E}$ can be accomplished by use of the gradient $\nabla\ell(\boldsymbol{\gamma})$ of $\ell(\boldsymbol{\gamma};\mathbf{x}) = \log p(\mathbf{x};\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$. Let

$$\hat{f}(\theta|\mathbf{x}) = \hat{p}(\mathbf{x}|\theta)\phi(\theta)/\hat{p}(\mathbf{x}) \tag{44}$$

be the estimated conditional density of θ_i, given $\mathbf{X}_i = \mathbf{x}$. Coordinate j of $\hat{\nabla}\ell(\mathbf{x}) = \nabla\ell(\hat{\boldsymbol{\gamma}};\mathbf{x})$ is

$$\hat{\ell}_j(\mathbf{x}) = \int [\theta(x_j - \hat{p}(1|\theta))]\hat{f}(\theta|\mathbf{x})d\theta, \tag{45}$$

and coordinate $q + j$ of $\nabla\hat{\ell}(\mathbf{x})$ is

$$\hat{\ell}_{q+j}(\mathbf{x}) = -\int [x_j - \hat{p}(1|\theta)]\hat{f}(\theta|\mathbf{x})d\theta \tag{46}$$

(Dempster, Laird, & Rubin, 1977).

Assume that $d(\mathbf{x})$ is not a linear function of $\nabla\ell(\boldsymbol{\gamma};\mathbf{x})$ for $\mathbf{x}$ in Γ. By a very similar argument to that required in the case of generalized residuals for log-linear models or for generalized residuals for indirectly observed log-linear models (Haberman, 1978, 1979), it follows that an estimate s of the standard deviation of $O - \hat{E}$ is $\hat{\tau}/n^{1/2}$, where $\hat{\tau}^2$ is the minimum of the weighted sum of squares

$$\sum_{\mathbf{x}\in\Gamma} \hat{p}(\mathbf{x})[d(\mathbf{x}) - \hat{E} - \boldsymbol{\alpha}'\hat{\nabla}\ell(\mathbf{x})]^2$$

for $2q$-dimensional vectors $\boldsymbol{\alpha}$. Let

$$\hat{\mathbf{W}} = \sum_{\mathbf{x}\in\Gamma} \hat{p}(\mathbf{x})\hat{\nabla}\ell(\mathbf{x})[\hat{\nabla}\ell(\mathbf{x})]' \tag{47}$$

and

$$\hat{\mathbf{u}} = \sum_{\mathbf{x}\in\Gamma} \hat{p}(\mathbf{x})[d(\mathbf{x}) - \hat{E}]\hat{\nabla}\ell(\mathbf{x}). \tag{48}$$

If

$$\hat{\boldsymbol{\alpha}} = \hat{\mathbf{W}}^{-1}\hat{\mathbf{u}}, \tag{49}$$

then

$$\hat{\tau}^2 = \sum_{\mathbf{x}\in\Gamma} \hat{p}(\mathbf{x})[d(\mathbf{x}) - \hat{E} - \hat{\boldsymbol{\alpha}}'\hat{\nabla}\ell(\mathbf{x})]^2. \tag{50}$$

Then $t = (O - \hat{E})/s$ has an approximate standard normal distribution if the model holds. Conditions for the approximation are similar to those for the Rasch model. The normal approximations are expected to be increasingly satisfactory if each ratio $d(\mathbf{x})/(n^{1/2}\hat{\tau})$ is small and each element of $(n\hat{\mathbf{W}})^{-1}$ is small.

In practice, s is often difficult to compute due to the large size of Γ. A simpler alternative replaces $\hat{p}(\mathbf{x})$ by the sample proportions $\bar{p}(\mathbf{x})$, so that $t_* = (O - \hat{E})/s_*$ is considered for $s_* = \hat{\tau}_*/n^{1/2}$. Here

$$\begin{aligned}\hat{\mathbf{W}}_* &= \sum_{\mathbf{x}\in\Gamma} \bar{p}(\mathbf{x})\hat{\nabla}\ell(\mathbf{x})[\hat{\nabla}\ell(\mathbf{x})]' \\ &= n^{-1}\sum_{i=1}^{n} \hat{\nabla}\ell(\mathbf{X}_i)[\hat{\nabla}\ell(\mathbf{X}_i)]', \end{aligned}\tag{51}$$

$$\begin{aligned}\hat{\mathbf{u}}_* &= \sum_{\mathbf{x}\in\Gamma} \bar{p}(\mathbf{x})[d(\mathbf{x}) - \hat{E}]\hat{\nabla}\ell(\mathbf{x}) \\ &= n^{-1}\sum_{i=1}^{n}[d(\mathbf{X}_i) - \hat{E}]\hat{\nabla}\ell(\mathbf{X}_i), \end{aligned}\tag{52}$$

$$\hat{\boldsymbol{\alpha}}_* = \hat{\mathbf{W}}_*^{-1}\hat{\mathbf{u}}_*, \tag{53}$$

and

$$\begin{aligned}\hat{\tau}_*^2 &= \sum_{\mathbf{x}\in\Gamma} \bar{p}(\mathbf{x})[d(\mathbf{x}) - \hat{E} - \hat{\boldsymbol{\alpha}}_*'\hat{\nabla}\ell(\mathbf{x})]^2 \\ &= n^{-1}\sum_{i=1}^{n}[d(\mathbf{X}_i) - \hat{E} - \hat{\boldsymbol{\alpha}}_*'\hat{\nabla}\ell(\mathbf{X}_i)]^2. \qquad (54)\end{aligned}$$

The statistic t_* also has an approximate standard normal distribution.

One simple application considers the cumulative distribution function for the raw total scores $S_i = S(\mathbf{X}_i)$. For an integer y from 0 to $q-1$, consider $d(\mathbf{x})$ equal to 1 for $S(\mathbf{x}) \leq y$ and to 0 otherwise. This case is of no interest for the conditional estimation in section 2, for the observed and fitted marginal distributions of S_i are the same; however, the 2PL model with a normal distribution for θ_i does not have this property. For the data under study, results are shown in Table 2. Values for y from 0 to 3 are omitted because no observed examinee has $S_i \leq 3$. As a consequence, t_* is not properly defined. In addition, given the sample size of 8,686 and standard rules for normal approximations of binomial distributions, normal approximations are questionable for y less than 8 or y greater than 41. These cases are also omitted. It is clearly true that the empirical distribution function of the S_i is not consistent with the 2PL model with a normal ability distribution. The discrepancies are not negligible, for the difference between empirical and fitted distribution functions is as large as 0.038. An obvious pattern exists, for extreme scores lead to negative generalized residuals, and scores near the center of the empirical distribution of the S_i lead to positive generalized residuals.

4 Conclusions

The methods for examination of residuals developed in this report are widely applicable. They can be employed with any item-response model such that maximum-likelihood estimates have normal approximations in large samples. Thus use is straightforward with partial-credit models, generalized partial-credit models, models with latent classes rather than continuous latent variables, and 1PL models with normal ability distributions. In principle, application to 3PL models should be straightforward; however,

Table 2

Cumulative Distribution Function of Raw Total Score for the Two-Parameter Logistic Model

Score	Observed CDF	Fitted CDF	Generalized residual
8	0.005	0.014	−12.444
9	0.010	0.021	−12.566
10	0.017	0.031	−11.902
11	0.026	0.044	−12.572
12	0.038	0.059	−13.723
13	0.057	0.078	−11.736
14	0.081	0.109	−10.364
15	0.113	0.127	−6.866
16	0.150	0.158	−3.353
17	0.191	0.192	−0.264
18	0.237	0.230	2.826
19	0.284	0.271	4.963
20	0.341	0.316	8.879
21	0.393	0.363	10.084
22	0.446	0.413	10.883
23	0.501	0.464	12.085
24	0.552	0.516	11.718
25	0.607	0.569	12.368
26	0.652	0.621	10.449
27	0.698	0.672	9.011
28	0.743	0.722	7.939
29	0.780	0.768	4.966
30	0.818	0.811	3.032
31	0.848	0.850	−0.531
32	0.875	0.884	−4.326
33	0.901	0.913	−7.063
34	0.923	0.938	−9.112
35	0.939	0.957	−12.084
36	0.956	0.972	−11.648
37	0.970	0.983	−10.675
38	0.980	0.991	−9.579
39	0.988	0.995	−8.283
40	0.993	0.998	−6.267
41	0.997	0.999	−4.778

Note. CDF = cumulative distribution function.

parameter estimates with very large asymptotic variances often have adverse impacts on implementation.

Numerous kinds of generalized residuals may be considered. For example, one may have sample statistics O equal to sample moments of total scores or average products of item scores and functions of total scores. One may also explore average products of distinct item scores to examine whether joint probabilities of correct pairs of item scores are properly predicted.

It should be emphasized that results for the specific data examined do not appear unusual. The methods used in this report have been applied to a variety of operational data with similar results. No item-response model appears to agree with any operational data. Whether the size of model errors is sufficiently great to adversely affect the practical application of item-response models is much less clear. This issue merits further study.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, *32*, 283–301.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, *34*, 42–54.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, *11*, 375–386.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*, 443–459.

Cochran, W. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, *10*, 417–451.

Cochran, W. G. (1955). A test of a linear function of the deviations between observed and expected numbers. *Journal of the American Statistical Association*, *50*, 377–397.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika*, *54*, 635–659.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). New York: Springer.

Haberman, S. J. (1976). Generalized residuals for log-linear models. In *Proceedings of the ninth International Biometrics Conference* (Vol. 1, pp. 104–172). Raleigh, NC: Biometrics Society.

Haberman, S. J. (1978). *Analysis of qualitative data: Vol. 1. Introductory topics.* New York: Academic Press.

Haberman, S. J. (1979). *Analysis of qualitative data: Vol. 2. New developments.* New York: Academic Press.

Haberman, S. J. (2004). *Maximum likelihood for the Rasch model for binary responses*

(ETS Research Rep. No. RR-04-20). Princeton, NJ: ETS.

Haberman, S. J. (2005). *Latent-class item response models* (ETS Research Rep. No. RR-05-28). Princeton, NJ: ETS.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute*, *22*, 719–748.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics*, *9*, 23–30.

Yates, F. (1948). The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, *35*, 176–181.