Some Goodness of Fit Measures for Cognitive Diagnosis Models

Presented by
Jimmy de la Torre
Department of Educational Psychology
Rutgers, The State University of New Jersey

- For a given data, several cognitive diagnosis models can be fitted
- The probability of correct classification depends on the fit of the models so the models need to be compared in terms of how well they fit the data
- Several goodness-of-fit measures will be covered in this presentation

- The measures can be classified as either at the item level or at the test level
- Three indices at the item level were computed by comparing the expected and observed characteristics of the marginal and pairwise joint distributions of the items
- Three other indices at the test level - Bayes factor, Aikake information criterion (AIC) and Bayesian information criterion (BIC) - were computed to provide global measures of the relative fits of the models

Item-level indices

- Compare expected and observed first and second moments
  - (1) proportion correct of each item
  - (2) correlation of each item-pair
  - (3) log-odds ratio of each item-pair
- Observed moments were computed using the data
- Expected moments were computed with model parameter estimates using simulation

- For the actual or simulated data, compute
  - (1) proportion correct for item *j*
  - (1) correlation of items *j* and *j'*
  - (2) log-odds ratio of items *j* and *j'*

$$\log\left[\frac{\Pr(Y_j = 1, Y_{j'} = 1) \times \Pr(Y_j = 0, Y_{j'} = 0)}{\Pr(Y_j = 1, Y_{j'} = 0) \times \Pr(Y_j = 0, Y_{j'} = 1)}\right]$$

- Compare observed and expected moments – large residuals indicate lack of fit

- How large is large? Need standard errors of the residuals
- Approximate standard errors of residuals by:
  - randomly selecting 100 draws of the structural parameters from the Markov chain output
  - obtaining expected moments for each draw
  - computing the standard deviation of moments across the 100 draws
- Standard errors can be used in evaluating the relative size of the residuals, not necessarily as a way of determining their significance

Fraction subtraction data: Proportion

		Residual			
	Observed	HO-NIDA		HO-DINA	
Item	Proportion	Estimate	S.E.	Estimate	S.E.
1	0.58	-0.17	0.01	0.00	0.01
2	0.53	0.09	0.01	0.00	0.01
3	0.79	0.05	0.01	0.00	0.01
4	0.40	0.02	0.01	0.00	0.01
5	0.65	-0.14	0.01	0.00	0.01
6	0.38	-0.06	0.01	0.00	0.01
7	0.47	0.03	0.01	0.00	0.01
8	0.73	0.02	0.01	0.00	0.01
9	0.72	0.11	0.01	0.00	0.01
10	0.45	0.07	0.01	0.00	0.01
11	0.70	0.08	0.01	0.00	0.01
12	0.43	-0.02	0.01	0.00	0.01
13	0.47	0.03	0.01	0.00	0.01
14	0.31	-0.07	0.01	0.00	0.01
15	0.39	-0.05	0.01	0.00	0.01

Mean Absolute Residuals: Second Moments

	HO-NIDA		HO-DINA	
Item	Corr.	Log(OR)	Corr.	Log(OR)
1	0.14	0.65	0.22	1.23
2	0.15	0.36	0.22	0.33
3	0.14	0.64	0.20	0.44
4	0.15	0.38	0.19	0.43
5	0.14	0.48	0.18	0.31
6	0.13	0.80	0.17	0.43
7	0.12	0.50	0.13	0.63
8	0.11	0.42	0.13	0.30
9	0.12	0.28	0.14	0.21
10	0.12	0.30	0.14	0.43
11	0.10	0.26	0.11	0.20
12	0.10	0.23	0.10	0.19
13	0.10	0.27	0.09	0.36
14	0.09	0.29	0.06	0.40
15	0.08	0.32	0.05	0.31

Global measures

- Instead of examining higher-order dependencies (e.g., item-triples), single global measures are computed for each model
- These measures can be used to compare competing non-nested models
- An important and common element of the three global measures is the conditional likelihood

- The conditional likelihood is given by

$$\Pr(\mathbf{Y} \mid \hat{\boldsymbol{\lambda}}, \hat{\mathbf{s}}, \hat{\boldsymbol{g}}) = \int\int \Pr(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \hat{\boldsymbol{\lambda}}, \hat{\mathbf{s}}, \hat{\boldsymbol{g}})\, d\boldsymbol{\theta}\; d\boldsymbol{\alpha}$$

- This can be approximated by quadrature nodes
- Bayes factor, AIC and BIC are functions of the conditional likelihood

Results for global measures

- Log of the Bayes factor: 407.22, in favor of the HO-DINA model
- Difference in AIC: 916.80, again in favor of the HO-DINA model
- Difference in BIC: 889.68, again in favor of the HO-DINA model
- Conclusion: HO-DINA provides a better fit for the fraction subtraction data

Summary

- Because model fit affects the probability of correctly classifying examinees in terms of their attributes, cognitive diagnosis models need to be evaluated in terms of their fit
- Indices are available to evaluate the model fit both at the item and test levels which have different usefulness

- Item level indices are useful in determining which specific items need closer inspection
- Test level indices are useful determining which models to select among competing models