Assessment of Model Fit in Diagnostic Classification Models

Section 5

Session Overview

- This section covers the varying ways of model fit can be assessed when using DCMs
  - These topics will apply to virtually any analysis of categorical data
- Model fit is used to help:
  - Determine if a model fits the data well enough in an absolute sense to use the examinee estimates
  - Select best model among competing models
- ECPE data will be used to illustrate model fit in practice

Section 5: Assessment of Model Fit

ASSESSMENT OF MODEL FIT

Assessing Model Fit

- There is no one best way to assess fit in DCMs
- Techniques typically used can be put into several general categories:
  - Absolute fit – global level
    - Model based hypothesis tests (if available)
    - Don't use entropy – misleading statistic
  - Absolute fit – item level
    - Fit for each item marginally
    - Fit for all pairs of items
  - Relative fit
    - Likelihood ratio tests for nested models
    - Information criteria for non-nested models
- Topics discussed here will mainly focus on fit statistics available in Mplus

The Big Picture of Model Fit

- Before using a DCM (or any statistical model), you must first check to see if it adequately represents the data
  - This is the test of absolute fit – how well your model fits the data you have in the absence of any competing models
- Should your model be shown to fit the data, you can safely use, interpret, and make inferences from the parameters of the model
  - Including examinee estimates
- If more than one model fits the data well, you can use relative fit statistics to decide which model is more appropriate
- If your model does not fit the data, you cannot use the results
  - Most parameters, their standard errors, and all hypothesis tests will be biased and misleading

EVALUATING ABSOLUTE FIT UNDER DCMS

Absolute Fit for Categorical Data Models

- When using categorical data, evaluating a model's absolute fit comes from a very familiar method: the classical Chi-Squared statistic comparing expected versus observed counts of examinees
  - The counts are across the entire set of items
- The Chi-Squared test still has the same requirements:
  - Each possible pattern must have been observed several times (some sources say 5, others 10)
- This makes the overall Chi-Squared test of model fit essentially useless for most data sets

Observed Vs. Expected Counts

- Observed counts: Number of people responding with each **possible** response pattern
  - For I binary items, there are 2^I possible response patterns
    - In the 28-item ECPE data, that means 268,435,456 patterns are possible
    - Our sample would have to be much larger than that just to find people that would have all patterns (try billions)
- Expected counts: come from diagnostic classification model probability times the sample size

$$N * P(\mathbf{X}_r = \mathbf{x}_r) = N * \sum_{c=1}^{C} \nu_c \prod_{i=1}^{I} {\pi_{ic}}^{x_{ir}} (1 - \pi_{ic})^{x_{ir}}$$

Obtaining Absolute Model Fit Statistics in Mplus

- Mplus provides absolute model fit statistics using the TECH10 option

```
OUTPUT:
TECH1 TECH5 TECH8 TECH10;
```

- Although it is impossible to typically get an accurate test of **global** model fit, other statistics can be obtained

- In any event, we will discuss what comes out of Mplus and how to use each portion

Overall Model Fit: Chi-Squared Test

- For small numbers of items (10-15), the traditional Chi-Squared test of model fit can be used
  - Test is invalid for too many items – sparse data

- For the 28 item test, the most recent version of Mplus does not provide this test, only a message:

```
THE MODEL ESTIMATION TERMINATED NORMALLY

    THE CHI-SQUARE TEST IS NOT COMPUTED BECAUSE THE FREQUENCY TABLE FOR THE
    LATENT CLASS INDICATOR MODEL PART IS TOO LARGE.
```

- Under the output section, though, Mplus will give expected and observed counts for all response patterns
  - In our data, there were 2690 patterns observed (sample was 2922)

Overall Chi-squared Test

- 2690 observed response patterns

RESPONSE PATTERN FREQUENCIES AND CHI-SQUARE CONTRIBUTIONS

Response Pattern	Frequency Observed	Frequency Estimated	Standardized Residual (z-score)	Chi-square Contribution Pearson	Chi-square Contribution Loglikelihood	Deleted
1	2.00	1.33	0.58	0.33	1.62	
2	1.00	0.40	0.94	0.87	1.81	
3	2.00	0.93	1.12	1.25	3.08	
4	78.00	14.27	16.91	284.69	265.00	
5	1.00	0.03	6.07	36.89	7.32	
6	3.00	1.31	1.47	2.16	4.95	
7	1.00	0.00	71.35	5091.35	17.07	DELETED
8	4.00	1.66	1.82	3.32	7.05	
9	1.00	0.00	110.17	12137.62	18.81	DELETED
10	1.00	0.01	10.72	114.83	9.52	DELETED
11	1.00	0.05	4.50	20.22	6.20	
12	1.00	0.09	2.95	8.67	4.72	
13	1.00	0.00	184.69	34110.34	20.87	DELETED
14	1.00	0.00	37.10	1376.48	14.46	DELETED
15	2.00	0.28	3.28	10.74	7.91	
16	1.00	0.00	59.11	3493.99	16.32	DELETED
17	1.00	0.00	14.58	212.47	10.74	DELETED
18	1.00	0.01	9.36	87.64	8.99	
19	5.00	2.79	1.32	1.74	5.82	
20	1.00	0.00	940.24	********	27.38	DELETED
21	1.00	0.41	0.92	0.84	1.77	
22	1.00	0.02	7.09	50.20	7.91	
23	1.00	0.00	120.75	14581.73	19.18	DELETED
24	1.00	0.00	390.30	********	23.87	DELETED
25	1.00	0.01	13.44	180.66	10.42	DELETED
26	1.00	0.00	36.69	1346.12	14.41	DELETED
27	1.00	0.00	681.59	********	26.10	DELETED
28	1.00	0.00	147.90	21875.39	19.99	DELETED
29	1.00	0.50	0.70	0.49	1.37	
30	1.00	0.00	52.53	2758.90	15.85	DELETED
31	1.00	0.08	3.34	11.13	5.14	
32	1.00	0.00	268.87	72293.42	22.38	DELETED
33	1.00	0.01	12.76	162.71	10.21	DELETED
34	9.00	3.95	2.54	6.44	14.80	
35	1.00	0.00	101.98	10400.06	18.50	DELETED
36	1.00	0.00	235.85	55627.47	21.85	DELETED
37	1.00	0.00	330.68	********	23.20	DELETED
38	8.00	3.96	2.03	4.13	11.26	
39	1.00	0.00	248.41	61707.14	22.06	DELETED
40	1.00	0.00	60.65	3678.76	16.42	DELETED
41	1.00	0.00	43.82	1920.02	15.12	DELETED
...						
2688	1.00	0.00	488.93	********	24.77	DELETED
2689	1.00	0.00	18.28	334.29	11.64	DELETED
2690	1.00	0.03	6.13	37.57	7.35	

THE TOTAL PEARSON CHI-SQUARE CONTRIBUTION FROM EMPTY CELLS IS 2703.49

What to do About Absolute Fit?

- Because no good tests of global model level absolute fit exist, we are forced to look at what we can observe:
  - Marginal item fit statistics
  - Bivariate item fit statistics
- Each of these statistics still uses the Chi-Squared statistic
  - The comparison of expected and observed happens now at the item level (for marginal item fit) and at the item-pair level (for bivariate item fit)
- The goal in this analysis is to determine which items (marginally) or pairs of items (bivariate) demonstrate poor fit – then fix the model (or remove the items)

Item Fit Statistics

- The TECH10 option reports a degree of misfit for each
  - Item individually (Univariate)
  - Pair of two items (Bivariate)
- Uses Chi-squared test for misfit
  - Values for each item are distributed as Chi-square with 1 df (for binary items)
- Misfitting items can be investigated
  - Q-matrix can be changed
  - Items can be removed

Tech 10 Item Fit Statistics: Univariate Fit

- Univariate fit attempts to determine if the model fits each item marginally
  - A limited information statistic – uses only a portion of the entire response pattern
- Not useful in DCMs
  - Model is for probability
  - Will always fit perfectly

UNIVARIATE MODEL FIT INFORMATION

Variable	Estimated Probabilities H1	H0	Standardized Residual (z-score)
X1			
Category 1	0.197	0.197	0.000
Category 2	0.803	0.803	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
X2			
Category 1	0.170	0.170	0.000
Category 2	0.830	0.830	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
X3			
Category 1	0.421	0.421	0.000
Category 2	0.579	0.579	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
X4			
Category 1	0.294	0.294	0.000
Category 2	0.706	0.706	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000

Univariate Log-Likelihood Chi-Square			0.000
X26			
Category 1	0.297	0.297	0.000
Category 2	0.703	0.703	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
X27			
Category 1	0.553	0.553	0.000
Category 2	0.447	0.447	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
X28			
Category 1	0.180	0.180	0.000
Category 2	0.820	0.820	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
Overall Univariate Pearson Chi-Square			0.000
Overall Univariate Log-Likelihood Chi-Square			0.000

Tech 10 Item Fit Statistics: Bivariate Fit

- Bivariate fit is an index of fit for a pair of items
- Compares observed data with frequency expected under DCM
  - Produces a 1-df Chi-Squared test
- Can help identify items that do not fit model
  - Rough approximation

BIVARIATE MODEL FIT INFORMATION

Variable	Variable	Estimated Probabilities H1	H0	Standardized Residual (z-score)
X1	X2			
Category 1	Category 1	0.045	0.042	0.956
Category 1	Category 2	0.152	0.156	-0.526
Category 2	Category 1	0.125	0.128	-0.571
Category 2	Category 2	0.678	0.674	0.407
Bivariate Pearson Chi-Square				1.448
Bivariate Log-Likelihood Chi-Square				1.428
X1	X3			
Category 1	Category 1	0.097	0.098	-0.235
Category 1	Category 2	0.101	0.099	0.234
Category 2	Category 1	0.324	0.322	0.150
Category 2	Category 2	0.479	0.480	-0.140
Bivariate Pearson Chi-Square				0.125
Bivariate Log-Likelihood Chi-Square				0.125
X1	X4			
Category 1	Category 1	0.074	0.069	0.987
Category 1	Category 2	0.124	0.128	-0.749
Category 2	Category 1	0.220	0.225	-0.600
Category 2	Category 2	0.582	0.578	0.507
Bivariate Pearson Chi-Square				1.783
Bivariate Log-Likelihood Chi-Square				1.771
X1	X5			
Category 1	Category 1	0.031	0.029	0.863
Category 1	Category 2	0.166	0.169	-0.385
...				
X26	X27			
Category 1	Category 1	0.178	0.175	0.405
Category 1	Category 2	0.119	0.122	-0.470
Category 2	Category 1	0.375	0.378	-0.317
Category 2	Category 2	0.327	0.324	0.329
Bivariate Pearson Chi-Square				0.464
Bivariate Log-Likelihood Chi-Square				0.465
X26	X28			
Category 1	Category 1	0.067	0.068	-0.320
Category 1	Category 2	0.231	0.229	0.192
Category 2	Category 1	0.114	0.112	0.255
Category 2	Category 2	0.589	0.590	-0.164
Bivariate Pearson Chi-Square				0.193
Bivariate Log-Likelihood Chi-Square				0.193
X27	X28			
Category 1	Category 1	0.124	0.112	2.103
Category 1	Category 2	0.430	0.442	-1.334
Category 2	Category 1	0.056	0.069	-2.618
Category 2	Category 2	0.390	0.378	1.366
Bivariate Pearson Chi-Square				12.467
Bivariate Log-Likelihood Chi-Square				12.745
Overall Bivariate Pearson Chi-Square				1052.514
Overall Bivariate Log-Likelihood Chi-Square				1046.630

Bivariate Item Fit

- H0: Under DCM (Model-expected)
  - The model (set of attribute profiles) accounts for bivariate dependencies between item responses
- H1: Observed Data
- Category 1: Incorrect Response
- Category 2: Correct Response

BIVARIATE MODEL FIT INFORMATION

Variable	Variable	Estimated Probabilities H1	H0	Standardized Residual (z-score)
X1	X2			
Category 1	Category 1	0.045	0.042	0.956
Category 1	Category 2	0.152	0.156	-0.526
Category 2	Category 1	0.125	0.128	-0.571
Category 2	Category 2	0.678	0.674	0.407
Bivariate Pearson Chi-Square				1.448
Bivariate Log-Likelihood Chi-Square				1.428
X1	X3			
Category 1	Category 1	0.097	0.098	-0.235
Category 1	Category 2	0.101	0.099	0.234
Category 2	Category 1	0.324	0.322	0.150
Category 2	Category 2	0.479	0.480	-0.140
Bivariate Pearson Chi-Square				0.125
Bivariate Log-Likelihood Chi-Square				0.125

Fail to reject assumption of independence (i.e., shows model fit for these item pairs)

How do you get the Chi-square Value?

	A	B	C	D	E	F	G	H	I
1	Items	H1	H0	N*H1	N*H0	N*H1-N*H0	(N*H1-N*H0)^2	(N*H1-N*H0)^2/(N*H0)	
2	1 & 2	0.045	0.042	131.49	122.724	8.766	76.842756	0.626142857	
3		0.152	0.156	444.144	455.832	-11.688	136.609344	0.299692308	
4		0.125	0.128	365.25	374.016	-8.766	76.842756	0.205453125	
5		0.678	0.674	1981.12	1969.43	11.688	136.609344	0.069364985	
6							Chi Square	1.200653275	
7							p-value	0.273191149	

BIVARIATE MODEL FIT INFORMATION

Variable	Variable	Estimated Probabilities H1	H0	Standardized Residual (z-score)
X1	X2			
Category 1	Category 1	0.045	0.042	0.956
Category 1	Category 2	0.152	0.156	-0.526
Category 2	Category 1	0.125	0.128	-0.571
Category 2	Category 2	0.678	0.674	0.407
Bivariate Pearson Chi-Square				1.448
Bivariate Log-Likelihood Chi-Square				1.428

Fail to reject assumption of independence (i.e., shows model fit for these item pairs)

How do you get the Chi-square Value?

	A	B	C	D	E	F	G	H	I
1	Items	H1	H0	N*H1	N*H0	N*H1-N*H0	(N*H1-N*H0)^2	(N*H1-N*H0)^2/(N*H0)	
14	26 & 27	0.124	0.112	362.328	327.264	35.064	1229.484096	3.756857143	
15		0.43	0.442	1256.46	1291.52	-35.064	1229.484096	0.951963801	
16		0.056	0.069	163.632	201.618	-37.986	1442.936196	7.156782609	
17		0.39	0.378	1139.58	1104.52	35.064	1229.484096	1.113142857	
18							Chi-square	12.97874641	
19							p-value	0.000315047	

X27	X28			
Category 1	Category 1	0.124	0.112	2.103
Category 1	Category 2	0.430	0.442	-1.334
Category 2	Category 1	0.056	0.069	-2.618
Category 2	Category 2	0.390	0.378	1.366
Bivariate Pearson Chi-Square				12.467
Bivariate Log-Likelihood Chi-Square				12.745
Overall Bivariate Pearson Chi-Square				1052.514
Overall Bivariate Log-Likelihood Chi-Square				1046.630

Reject assumption of independence (i.e., shows model does not fit for this item pair)

Examining Bivariate Item Misfit

- From the output, we find that misfitting items often show up in multiple pairings:
  - Item 13: 13 significant misfitting pairs
  - Item 15: 12 significant misfitting pairs
  - Item 4: 11 significant misfitting pairs
  - Item 9: 11 significant misfitting pairs

- Overall, we found 90 pairs with significant values of misfit
  - We had a total of $\frac{28(28-1)}{2} = 378$ pairs of items to examine
- Using a Type-I error rate of 0.05 (significant Chi Square > 3.84), we would have expected to find 19 pairs significant by chance
  - Conclusion: our model doesn't fit well

Next Steps: Correcting Item Misfit

- Misfitting item pairs can happen for a number of reasons:
  - Incorrect Q-matrix for either item
  - Both items measure another attribute in the Q-matrix
  - Both items measure another attribute not in the Q-matrix

- The process of modification of the Q-matrix must be guided by substantive theory
  - Estimation must be conducted again

- Another option: omit item altogether
  - As the Rasch people do!
  - Models with different numbers of items cannot be compared using the relative fit statistics we are about to discuss

ENTROPY: A USELESS MEASURE OF MODEL FIT

Overall Model Fit: (Relative) Entropy

- The entropy of a model is a measure of classification uncertainty
  - It is an absolute fit statistic
- Mplus reports relative entropy
  - Value of 1.00 means all respondents classified with complete certainty (good fit)
  - Value of 0.00 means all respondents classified with equal probabilities for all classes (poor fit)
- ECPE (relative) entropy: 0.672
  - Hard to interpret by itself
- Of note: badly misfitting models (see DINA) will have very high entropy
  - Therefore – don't use entropy unless you have checked absolute fit

RELATIVE TESTS OF MODEL FIT

Relative Tests of Model Fit

- Once absolute fit has been established for more than one model – the next step is to use relative model fit statistics to choose the best/most parsimonious model
- Generally, relative model fit relates to comparing nested models – for which we can use likelihood ratio tests
  - LR tests sometimes have issues with parameters set to boundaries – more on this shortly
  - LR tests rely on the estimation procedure using marginal maximum likelihood
- For models that are not nested, we use information criteria to compare fit
  - Based on log-likelihoods, but corrected for model parsimony

Likelihood Ratio Tests

- Comparisons of nested DCMs can be conducted using a Likelihood Ratio test (LR test; also called a deviance test)
- Examples of **nested models** (single item or multiple items) where LR tests can be used:
  - LCDM v. DINA/DINO/rRUM/NIDA/NIDO/CRUM
  - Tests of LCDM item parameters: addition/removal of terms (using same Q-matrix)
  - Tests of LCDM structural model parameters: addition/removal of terms (using same Q-matrix)
  - Q-matrices of different sizes (can be tricky, but can be done)
  - Models with continuous latent traits/polytomous attributes
- Examples of models where LR tests CANNOT be used:
  - Comparisons of tests with differing numbers of items
- LR tests are preferred for model comparison – so if you have the ability to use one, do so (don't pick information criteria)

How Likelihood Ratio Tests Work

- Likelihood ratio tests work by comparing two nested models – both must be estimated:
  - H0 (null model): the simpler model
  - H1 (alternative model): the more complex model
- The LR test statistic comes from the log-likelihood of both models (ℓ_{H1} from H1, ℓ_{H0} from H0):

$$LR = -2(\ell_{H0} - \ell_{H1})$$

- The LR test is compared to a χ^2 distribution where the degrees of freedom is:

$$df_{LR} = df_{H1} - df_{H0}$$

- The hypothesis test is for the parameters under constraint in the H1 model

LR Test Example: Removal of Structural Model 3-Way Interaction

- To demonstrate the LR test, we will compare two models from our previous section: the full ECPE analysis and the ECPE analysis without the 3-way interaction
  - Note: we are skipping the part where both models must exhibit good absolute fit

- LR Test Hypotheses:

H0 - Null model (simpler): $\gamma_{3,(1,2,3)} = 0$

H1 – Alternative model: $\gamma_{3,(1,2,3)} \neq 0$

LR Test Example

- From Mplus output of both models:

H0 (reduced model):

Number of Free Parameters	80
Loglikelihood	
H0 Value	-42739.827
H0 Scaling Correction Factor for MLR	1.054

H1 (full model):

Number of Free Parameters	81
Loglikelihood	
H0 Value	-42739.712
H0 Scaling Correction Factor for MLR	1.073

$$LR = -2(-42{,}739.827 - -42{,}739.712) = 0.23$$
$$df_{LR} = 81 - 80 = 1$$
$$p = .632$$

- The LR test p-value can be obtained from Excel "=chidist(0.23,1)"
- Conclusion: fail to reject H0 – the simpler model is retained

Issues in LR Tests

- When using a LR test, you must be sure that the parameters of the simpler model are not on their boundary
  - Means typically Chi-Square test is invalid
- Cases where this occurs:
  - Evaluating main effects (cannot be less than zero)
  - Comparing the DINA/DINO/NIDA/NIDO with the LCDM
    - Involves main effects
  - Evaluating attribute hierarchies
- The correct reference distribution is a mixture of Chi-Squares, which is typically obtained through simulation for tests involving more than one parameter

Relative Model Fit: Information Criteria

- Used when comparing between two models that are not nested
  - Can happen – but usually LR tests can be used

Information Criteria

Number of Free Parameters	81
Akaike (AIC)	85641.425
Bayesian (BIC)	86125.807
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	85868.440

- Mplus reports:
  - AIC and BIC
  - Sample size adjusted BIC
- All can be used
  - Smallest value is best

Session 6: Assessment of Model Fit

CONCLUDING REMARKS

Concluding Remarks: Model Fit

- Assessment of model fit in DCMs is currently a difficult task
  - Easily accessible options are limited
  - Can quickly find options that take longer to assess fit than to estimate model
  - Mplus options are adequate for initial screening
- DCMs share this problem with IRT models
  - General categorical data analyses
- Other model fit options are available and forthcoming
  - Based on limited information
  - Need further testing