Assessment of Model Fit

Lecture #6
ICPSR Item Response Theory Workshop

Lecture Overview

• An overview of model fit for IRT Models using Mplus

• 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

• Fraction subtraction data will be used to illustrate model fit in practice
ASSESSMENT OF MODEL FIT

Assessing Model Fit

• There is no one best way to assess fit in IRT Models

• Techniques typically used can be put into several general categories:
  ➢ Absolute fit
    • Model based hypothesis tests (if available)
  ➢ Relative fit
    • Information criteria
  ➢ Item fit
    • Univariate
    • Bivariate

• Topics discussed here will mainly focus on fit statistics available in Mplus
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

• Mplus gives this automatically
  ➢ Omits when data are sparse
    ➢ Can omit extreme cells from an analysis
      • Misleading

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
Relative Model Fit: Information Criteria

- Used when comparing between two models
  - 1PL v. 2PL
- Mplus reports:
  - AIC and BIC
  - Sample size adjusted BIC
- All can be used
  - Smallest value is best
- Here, 2PL Model is Preferred using AIC/BIC

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Free Parameters</th>
<th>Loglikelihood</th>
<th>H0 Value</th>
<th>Information Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1PL</td>
<td>21</td>
<td></td>
<td>-4797.178</td>
<td>Akaike (AIC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bayesian (BIC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sample-Size Adjusted BIC (n^* = (n + 2) / 24)</td>
</tr>
<tr>
<td>2PL</td>
<td>40</td>
<td></td>
<td>-4640.159</td>
<td>Akaike (AIC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bayesian (BIC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sample-Size Adjusted BIC (n^* = (n + 2) / 24)</td>
</tr>
</tbody>
</table>

Chi-Squared (Deviance) Test

- The 1PL and 2PL are nested models
  - Can use deviance test to statistically test for fit
- Change in -2*loglikelihood:
  - \(-2(-4797.178 - -4640.159) = 314.038\)
- Change in DF:
  - 40-21 = 19
- Chi-Square p-value < 0.0001
- Conclusion: 2PL is preferred statistically

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Free Parameters</th>
<th>Loglikelihood</th>
<th>H0 Value</th>
<th>Information Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1PL</td>
<td>21</td>
<td></td>
<td>-4797.178</td>
<td>Akaike (AIC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bayesian (BIC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sample-Size Adjusted BIC (n^* = (n + 2) / 24)</td>
</tr>
<tr>
<td>2PL</td>
<td>40</td>
<td></td>
<td>-4640.159</td>
<td>Akaike (AIC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bayesian (BIC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sample-Size Adjusted BIC (n^* = (n + 2) / 24)</td>
</tr>
</tbody>
</table>
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
  - Items can be removed
  - Multidimensional model may be used

Item Fit Statistics: Univariate Fit

- Univariate fit attempts to determine if the model fits each item marginally
  - Limited information statistic

- Not *that* useful in IRT
  - Scale of fit is usually small
  - Most items “fit”

- H1: Observed Probability

- H0: IRT Model Prediction

- Chi-Square critical values
  - (0.05) = 3.84
  - (0.01) = 6.63
Item Fit Statistics: Bivariate Fit

• Bivariate fit is an index of fit for a pair of items

• Compares observed data with frequency expected under IRT model
  ➢ Produces a 1-df Chi-Squared test for binary items

• Can help identify items that do not fit model
  ➢ Rough approximation

Table: Bivariate Model Fit Information

<table>
<thead>
<tr>
<th>Variable</th>
<th>Expected Residual (z-score)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fraction Subtraction Results: Model Fit

• Univariate model fit
  ➢ Compares model predicted and observed frequencies of responses for all items marginally
  ➢ Of 20 items, none had p-values less than 0.01

• Bivariate model fit
  ➢ Compares model predicted and observed frequencies of responses for all pairs of items
  ➢ Of 190 item pairs 35 had p-values less than 0.01
  ➢ Items most indicated
    • Item 4 (8 pairs): $\frac{1}{3} - \frac{2}{3}$
    • Item 6 (7 pairs): $\frac{1}{7} - \frac{4}{7}$
    • Item 20 (7 pairs): $\frac{1}{3} - \frac{5}{3}$

• Indicates some items are not fit well by model
  ➢ We will ignore this and continue with analysis as example
Residual Analysis

- Assess the accuracy of model predictions versus actual data
  - Residual = difference between observed proportion and predicted probability:
    \[ r_i(\theta) = P_i(\theta) - E[P_i(\theta)] \]
- \( P_i(\theta) \) = observed proportion correct for a given \( \theta \) level
- \( E[P_i(\theta)] \) = expected proportion correct (i.e., probability from the IRT model)
Residual Analysis

\[ r_i(\theta) = P_i(\theta) - E[P_i(\theta)] \]

White = \( P_i(\theta) \)
Black = \( E[P_i(\theta)] \)

Standardized Residuals

- Raw residuals do not take into account the error associated with the expected proportion correct, so we standardize each by dividing by its standard error:

\[
SE(E[P_i(\theta)]) = \sqrt{E[P_i(\theta)]E[1 - P_i(\theta)] / N(\theta)}
\]
Standardized Residuals

\[ SR_i(\theta) = \frac{P_i(\theta) - E[P_i(\theta)]}{\sqrt{E[P_i(\theta)]E[1 - P_i(\theta)]/N}} \]

SR values should be homoscedastic for each item and follow an approximately standard normal distribution across all items of the test.
SR values are homoscedastic for this item when fit by a 3-PL model, but systematic errors are present for the 1- and 2-PL models.

Test-level Fit

- Similar to the comparison done for individual items, but instead we compare Expected Proportion Correct (TCC) to observed proportion correct (raw score/N)

- SRs across the test should be homoscedastic and follow an approximate normal distribution
Across all items, SR values are approximately normally distributed when fit by a 3-PL model, but more uniform for the 1- and 2-PL models.

Significance Testing
Q1 chi-square (Yen, 1981)

\[ Q1_j = \sum_{i=1}^{m} SR_{ij}^2 \quad Q1 \sim \chi^2 \quad df = m - p \]

\[ m = \# \text{ of quadrature points} \]
\[ p = \# \text{ of item parameters} \]
Chi-square test

• Standardized residuals are essentially prediction errors that have been turned into z-scores

• The sum of squared z-scores follow a Chi-square distribution
  ➢ Much like Sums of Squares and variances follow a Chi-square distribution in ANOVA
As the degrees of freedom increase, the sum of squared standardized residuals are less likely to be equal to zero.

The expected value for $\text{SUM}(SR^2)$ moves to the right as the distribution approaches normality.

**Goodness-of-Fit**

- This is what we call “goodness-of-fit”:

- We hope that the chi-square test will NOT be significant
  - This indicates that the differences between observed and expected is small
  - Significant differences would mean that observed proportions are far from what the model predicted
Significance Testing in BILOG and PARSSCALE

- The goodness of fit information contained in BILOG and PARSSCALE use the Chi-square test described in the previous slides.

> These values can be found for all items

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Intercept</th>
<th>Slope</th>
<th>Threshold</th>
<th>Loading</th>
<th>Asymptote</th>
<th>ChiSq</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATH01</td>
<td>1.041</td>
<td>0.651</td>
<td>-1.599</td>
<td>0.545</td>
<td>0.186</td>
<td>29.0</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>0.107*</td>
<td>0.082*</td>
<td>0.242*</td>
<td>0.069*</td>
<td>0.084*</td>
<td>(0.0007)</td>
<td></td>
</tr>
<tr>
<td>MATH02</td>
<td>2.230</td>
<td>0.600</td>
<td>-3.717</td>
<td>0.514</td>
<td>0.199</td>
<td>9.5</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>0.165*</td>
<td>0.114*</td>
<td>0.610*</td>
<td>0.098*</td>
<td>0.089*</td>
<td>(0.0920)</td>
<td></td>
</tr>
<tr>
<td>MATH03</td>
<td>0.428</td>
<td>0.693</td>
<td>-0.618</td>
<td>0.569</td>
<td>0.159</td>
<td>63.2</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td>0.106*</td>
<td>0.084*</td>
<td>0.130*</td>
<td>0.069*</td>
<td>0.071*</td>
<td>(0.0000)</td>
<td></td>
</tr>
<tr>
<td>MATH04</td>
<td>-0.601</td>
<td>1.391</td>
<td>0.432</td>
<td>0.812</td>
<td>0.217</td>
<td>10.7</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>0.216*</td>
<td>0.268*</td>
<td>0.095*</td>
<td>0.156*</td>
<td>0.040*</td>
<td>(0.2204)</td>
<td></td>
</tr>
</tbody>
</table>

The problem of sample size

- Statistical tests of model-data fit present an interesting duality:

  - Due to sensitivity to sample size, almost any departure of data from the model results in rejecting H0

  - For small samples, model-data misfit can be overlooked, and SEs for item parameters are large
Concluding Remarks: Model Fit

• Assessment of model fit in IRT Models 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

• IRT models share this problem general categorical data analysis techniques

• Other model fit options are available and forthcoming