Test Construction

NCME 2009
Workshop

Introduction

- Now that we have discussed the basic concepts of the models, their estimation, and final fit evaluation we want to consider how this information can be used to construct an optimal test.

- Specifically, if one were to know the item parameters, how might one construct or refine a test as to use only the "good" items

- Therefore, we will now discuss methods to quantify a good item, which will help with test construction
  - Good in this case will be an item with high "discrimination"

Introduction

- In doing this we will first take an approach that is descriptive
  - Related to item discrimination from Classical Test Theory (CTT)
- Then, I will briefly discuss a method based on the Kullback-Leibler Information
  - Closely related to the goals of Item Response Theory (IRT)

CTT Discrimination

- One method of measuring an item discrimination using Classical Test Theory is to compute the point biserial correlation
  - High positive values are good
- However, as an alternative we may consider a basic comparison of probabilities
  - Those who have performed well on the test (Maybe top 25%)
  - Those who have performed poorly on the test (Maybe Lower 25%)

CTT Discrimination

- If the probability of answering the item right is very different for the two groups ($d_i = p_u - p_l$) then the item discriminates well

  where:

  - Let p_u denote the proportion of correct responses to an item for respondents from the upper tail of the total score distribution

  - Let p_l denote the proportion of correct responses to an item for respondents from the lower tail of the total score distribution

General DCM Item Discrimination

- This idea of comparing two probabilities can be used in DCMs.
- A person who has mastered all required attributes is expected to perform well
- A person who has not mastered any of the required attributes is expect to perform poorly

General DCM Item Discrimination

- A general definition of item discrimination is:

$$d_i = p_{\alpha_h} - p_{\alpha_l}$$

Probability of a correct response for masters of all measured attributes

Probability of a correct response for nonmasters of all measured attributes

Example

- Imagine that we have fit the LCDM for the item 4+1-2=?

- In this case we will get a parameter for the
  - Intercept
  - Main Effect of Addition
  - Main Effect of Subtraction
  - Interaction between Addition and Subtraction

$$P(X_{ij}=1|\alpha)=\frac{\exp(\lambda_{i,0}+\lambda_{i,(add)}\alpha_{j,add}+\lambda_{i,(sub)}\alpha_{j,sub}+\lambda_{i,(add,sub)}\alpha_{j,add}\alpha_{j,sub})}{1+\exp(\lambda_{i,0}+\lambda_{i,(add)}\alpha_{j,add}+\lambda_{i,(sub)}\alpha_{j,sub}+\lambda_{i,(add,sub)}\alpha_{j,add}\alpha_{j,sub})}$$

Example

- Including a possible set of values

$$P(X_{ij}=1|\alpha)=\frac{\exp(-1+1*\alpha_{j,add}+.5*\alpha_{j,sub}+1*\alpha_{j,add}\alpha_{j,sub})}{1+\exp(-1+1*\alpha_{j,add}+.5*\alpha_{j,sub}+1*\alpha_{j,add}\alpha_{j,sub})}$$

- And so:

$$p_{\alpha_h}=\frac{\exp(-1+1+.5+1)}{1+\exp(-1+1+.5+1)}=0.82$$

$$p_{\alpha_l}=\frac{\exp(-1)}{1+\exp(-1)}=0.27$$

A difference of 0.55

Attribute Discrimination

- Although the general measure of discrimination can be useful, it assumes that all attributes are being equally measured

- As an alternative we may be interested in quantifying the discrimination of an item for a particular attribute

- Here we can use the same concept, but now we are comparing the probability of a correct response given the attribute is mastered to the probability of a correct response assuming the attribute is not mastered

α_h

Attribute Discrimination

- In this case, the change in probability may actually depend on mastery of additional attributes

- Using the previous example focusing on Addition:

Frame for Comparison	(Add, Sub)	(Add, Sub)
1	(1, 1)	(0, 1)
2	(1, 1)	(0, 0)
3	(1, 0)	(0, 0)
4	(1, 0)	(0, 1)

Attribute Discrimination

- One method is to simply pick the comparison that will maximize the discrimination index

- We must also assume that all other relevant attributes are fixed to be the same in the comparison

Frame for Comparison	(Add, Sub)	(Add, Sub)
1	(1, 1)	(0, 1)
~~2~~	~~(1, 1)~~	~~(0, 0)~~
3	(1, 0)	(0, 0)
~~4~~	~~(1, 0)~~	~~(0, 1)~~

Attribute Discrimination

$$P(X=1|\alpha=\{1,1\})=\frac{\exp(-1+1+.5+1)}{1+\exp(-1+1+.5+1)}=0.82$$

$$P(X=1|\alpha=\{0,1\})=\frac{\exp(-1+.5)}{1+\exp(-1+.5)}=0.38$$

$$P(X=1|\alpha=\{1,1\})=\frac{\exp(-1+1)}{1+\exp(-1+1)}=0.50$$

$$P(X=1|\alpha=\{0,0\})=\frac{\exp(-1)}{1+\exp(-1)}=0.27$$

Attribute Discrimination

- So to complete the example for addition the two comparisons we have are :

Frame for Comparison	(Add, Sub)	(Add, Sub)	
1	(1, 1)	(0, 1)	d_j=.82-.38=**.44**
~~2~~	~~(1, 1)~~	~~(0, 0)~~	
3	(1, 0)	(0, 0)	d_j=.50-.27=.23
~~4~~	~~(1, 0)~~	~~(0, 1)~~	

Attribute Discrimination

- One challenge of attribute discrimination is that multiple comparisons must be made
- However, if we are using specific models, this definition allows us to develop a set of general guidelines
- Specifically, for the LCDM you can simply use the magnitude of the weights associated with each attribute as a quick guideline
- In addition, for models such as the DINA, DINO, RRUM, and the Compensatory RUM we can use the following…

Attribute Discrimination

Model	Global Item Discrimination	Attribute-specific Item Discrimination	A "good" item is one where...
DINA	$d_{i,DINA} = (1 - s_i) - g_i$	$d_{ia,DINA} = (1 - s_i) - g_i$	s_i and g_i are low
NC-RUM	$d_{i,NC-RUM} = \pi_i^* - \pi_i^* \prod_{a=1}^{A} r_{ia}^{*q_{ia}}$	$d_{ia,NC-RUM} = \pi_i^* - \pi_i^* r_{ia}^*$	π_i^* is high and r_{ia}^*s are low
DINO	$d_{i,DINO} = (1 - s_i) - g_i$	$d_{ia,DINO} = (1 - s_i) - g_i$	s_i and g_i are low
C-RUM	$d_{i,C-RUM} = \frac{\exp\left(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1(a)} q_{ia}\right)}{1 + \exp\left(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1(a)} q_{ia}\right)} - \frac{\exp(\lambda_{i,0})}{1 + \exp(\lambda_{i,0})}$	$d_{ia,C-RUM} = \frac{\exp(\lambda_{i,0} + \lambda_{i,1(a)})}{1 + \exp(\lambda_{i,0} + \lambda_{i,1(a)})} - \frac{\exp(\lambda_{i,0})}{1 + \exp(\lambda_{i,0})}$	$\lambda_{i,0}$ is low and the remaining $\lambda_{i,1,(a)}$ are high

- Rupp, Temple, and Henson (Forthcoming)

IRT Approach

- Now we want to move to briefly introduce a concept that is similar to what you may expect in the context of IRT.

- That is, a typical method of item selection and determining an items value is based on Fisher's Information
  - Largely because of its relationship with Standard Error of Estimate

- Requires a continuous variable

DCMs

- So, our definition of a "good" test must be slightly changed.
- We will need to define what is meant by a good test because measurement error does not mean quite the same thing with latent classes.
- A "good" test is one that correctly classifies examinees.
  - Correctly estimates examinees' profiles.

Objective

- It is our goal to define an index or set of indices that:
  - Relate to correct classification rates.
  - Have similar properties as in IRT.
    - Uses all of the relevant information.
    - Have a meaningful interpretation.
    - Defined for the item and the test.
  - Are additive (the test index should equal the sum of item index).

Discrimination Indices

- We will define a set of indices that have these characteristics.
  - Kullback-Leibler Information.
  - Test discrimination index for DCMs (C_j).

Kullback-Leibler Information

- The Kullback-Leibler information, $\delta[f, g]$, is most commonly described as a distance measure.
  - Specifically, the "distance" between the two probability distributions $f(X)$ and $g(X)$.

$$\delta(f,g) = E_f\left[\log\left[\frac{f(X)}{g(X)}\right]\right]$$

Kullback-Leibler Characteristics

- Not quite a distance.
  - It is not symmetric.
  - Does not satisfy the triangle inequality.
- But, the higher the value the easier it is to discriminate between the two distributions.
- If the distributions are the same then $\delta(f, g)=0$.

K-L for DCMs

- For DCMs we can start by thinking about any two skill patterns $\boldsymbol{\alpha_u}$ and $\boldsymbol{\alpha_v}$.

- We define:

$$f(x) = P(X_j \mid \alpha_u)$$

$$g(x) = P(X_j \mid \alpha_v)$$

- So:

$$\delta_j[\alpha_u, \alpha_v] = E_{\alpha_u}\left[\log\left[\frac{P(X_j \mid \alpha_u)}{P(X_j \mid \alpha_v)}\right]\right]$$

K-L for DCMs

- The K-L defined in this way will measure the degree to which the distributions differ.
  - This also is an indication of how well we can discriminate between skill pattern u and skill pattern v.
- Also, based on its definition, the test K-L comparing u to v is simply the sum of all item K-L for these two skill patterns.

Item Discrimination

- However, there are $2^k(2^k-1)$ possible pairs of comparison

- So one possible method is to summarize these in a weighted average based on how distinct the attribute patterns are

$$C_j = \frac{1}{\sum_{u \neq v} h(\alpha_u, \alpha_v)^{-1}} \sum_{u \neq v} h(\alpha_u, \alpha_v)^{-1} D_{juv}$$

Item Discrimination

- In addition, any method of summarizing these comparisons would work well as a possible index of discrimination

- Although I will not go in to detail here, there are attribute discrimination indices.

Summary

- In defining these new indices, we are able to determine the value of each item relative to all items being considered.

- In using this, we can refine, and construct "good" tests from a prespecified set of items.