Introduction and Psychometric Overview

PRE 932: Diagnostic Testing
August 27th, 2014
Lecture #1
Lecture Topics

• A brief history of measurement

• Item response theory -- and how it has come to be used in the psychometric community

• What IRT is...and isn’t
  ➢ Comparisons with other measurement models you may have heard of or used before

• A brief primer on statistics, logits, and other math things
IRT is a Part of a Broader Field: Test Theory

• Test theory :: Psychometric Theory
  ➢ A general collection of statistical techniques used for evaluating and developing psychological tests
  ➢ One of three dominant measurement paradigms
    • All three are interrelated

• Although IRT was developed because of the needs of certain psychological tests, its use has become much more widespread (e.g., used in Political Science)

• Now a part of a broader set of statistical techniques
  ➢ Generalized linear mixed models
What is a *Latent Trait*?

- Latent trait: An unobservable ability or characteristic
  - e.g., “intelligence”, “extroversion”, or “political idealization”

- A person’s latent trait(s) are estimated (measured) using a *measurement model*
  - Measurement model: A statistical model linking the unobserved latent trait with the observed outcome
    - In social/education research outcomes are generally test items
      - We will use the term item throughout

- Latent traits are measured with multiple observed items
  - Utilize common (co)variance among items
A (Very) Brief History of Test Theory

• Modern beginnings date to mid 19th century
  ➢ Measurement of intelligence

• 1904 brought about two seminal papers by Charles Spearman
  ➢ One showed how to estimate the amount of error in test scores
    ♦ Led to field of Classical Test Theory (CTT)
  ➢ One showed how measure a single trait from a test
    ♦ Led to field of factor analysis
    ♦ Modern versions feature measurement models under the name of Confirmatory Factor Analysis (CFA)
Development of the Field of Test Theory

- Motivated by problems in education and psychology
  - Education :: measuring intelligence or achievement
  - Psychology :: understanding structure of traits

- Early theory developed prior to computers
  - Work prior to the 1960s relied on approximations
  - IRT was developed largely in the 1960s and 1970s

- Mathematicians and statisticians have advanced the field in recent years
  - Brought rigor and validity to approaches
Measurement Models

• Measurement models can be divided into two families of models based on response format alone:
  - Continuous responses :: Confirmatory Factor Models
  - Categorical responses :: Item Response Models

• Both of these families fall under a larger framework: Generalized Linear Latent and Mixed Models
  - Provide measurement models for other types of responses

• Other relevant families (not covered in this workshop):
  - Structural Equation Models :: provides estimates of correlations amongst latent variables in measurement models
  - Path Analysis :: simultaneous regression amongst multiple observed variables
Differences Among Measurement Models

• Fundamental difference is in unit of analysis
  - **Classical Test Theory (CTT)** :: unit of analysis is the *entire test*
    - Sum of items = latent trait estimate
    - Positives: Can always be done; No need for advanced computing
    - Negatives: Restrictive assumptions; limited generalizability
  - **CFA and IRT** :: unit of analysis is the *item*
    - Model how item response relates to latent trait
    - Different models for different types of item response formats
    - Provides a framework for testing adequacy of measurement models

• Each family of models has a different name for the trait:
  - **CTT** :: True Score (T)
  - **CFA** :: Factor Score (F)
  - **IRT** :: Ability (commonly); Theta (θ)
Classical Test Theory Basics

• In CTT, the **test** is the unit of analysis:
  \[ Y_{total} = T + e \]
  - **True score** \( T \): best estimate of “latent trait”, mean over infinite replications of the test
  - **Error** \( e \): mean of zero, uncorrelated with \( T \)

• Variance of test scores: \( \sigma_Y^2 = \sigma_T^2 + \sigma_e^2 \)

• Goal is to quantify **reliability** :: proportion of test variance accounted for by true score variance:
  \[ \rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2} \]

• Items are assumed to be exchangeable (all count the same)
  - More items means higher reliability, regardless of type
More Classical Test Theory

• Error is a unitary construct in CTT
  ➢ Error variance has been quantified in various ways
  ➢ Goal is to reduce error variance as much as possible
    ♦ Standardization of testing conditions (reduces confounds)
    ♦ Aggregation of additional items (errors should cancel out)
  ➢ Items are exchangeable

• Followed by *generalizability theory* to decompose error
  ➢ e.g., rater variance, person variance, time variance...
Even More Classical Test Theory

• Brief history of solutions for quantifying reliability:
  ➢ 1904: Spearman:: from alternate forms or test-retest
  ➢ 1945: Guttman:: from the relations between the items within a test (i.e., coefficient alpha)
  ➢ 1951: Cronbach further developed Guttman’s work “Cronbach’s alpha”
    • Cronbach’s work further elaborated into generalizability theory
  ➢ 1950: Gulliksen classic text for CTT
    • See also Nunnally’s texts from the 1970’s - 1990’s

• Around that point, psychometrics started to shift to focus on the item
  ➢ Although the item had been investigated for years in another framework (CFA)
Developing Statistical Models for Test Data

At this point we will diverge from psychometric history and review some basic statistical models that will help in developing CFA and, ultimately, IRT

- In sum: we need to discuss linear regression

Imagine that you have:

- A **continuous** outcome variable: \( Y \)
- A **continuous** predictor variable: \( X \)

You wish to examine the relationship between \( X \) and \( Y \), using values of \( X \) to predict values of \( Y \)
Linear Regression (Both X and Y Observed)

- The prediction of Y is done using a linear regression:

\[ Y = \beta_0 + \beta_1 X + e \]

\( \beta_0 \) is the intercept (where the line crosses the Y axis)

\( \beta_1 \) is the slope (the increase in Y for a one unit increase in X)

\( e \) is the error (or residual), with estimated error variance \( \sigma_e^2 \)
Confirmatory Factor Analysis (CFA) Models

- Main idea of CFA:: Build a measurement model for response variables that measure the same trait
  - **CFA = Linear regression model** predicting each continuous observed outcome variable (item, subscale) from a latent trait predictor variable(s)

\[ Y_{si} = \mu_i + \lambda_i F_s + e_{si} \]

- \( s = \text{subject} \)
- \( i = \text{item} \)
- \( \mu_i \) is the item intercept
- \( \lambda_i \) is the item slope (factor loading)
- \( e_{si} \) is the error for the item and subject
- \( Y_{si} \) is the item response (**continuous**) to item \( i \) for subject \( s \)
Confirmatory Factor Analysis (CFA) Models

• CFA differs from exploratory factor analysis (which is not a model if conducted as it typically is with principal components-based methods):
  - Number and content of factors is decided a priori
  - Alternative models are comparable and testable

• Uses of confirmatory factor analysis models:
  - Analyze relationships among subscales that have normal, continuous distributions (or “incorrectly” to analyze item-level data)
  - Provide comparability across persons, items, and occasions
Factor Analysis (Y Observed; F latent)

- The prediction of Y is done using a linear regression:

\[ Y_{si} = \mu_i + \lambda_i F_s + e_{si} \]

- \( \mu_i \) is the item intercept (where the line crosses the Y axis)
- \( \lambda_i \) is the item’s slope (the increase in Y for a one unit increase in F)
- \( e_{si} \) is the error (or residual), with estimated error variance \( \psi_i^2 \)
Confirmatory Factor Analysis (CFA)

• Dimensionality is assumed to be known
  ➢ Local Independence is assumed → conditional on the number of dimensions in the model
  • Errors are independent after controlling for factor(s)

• CFA is a **linear model** :: a one-unit change in latent trait/factor F has same increase in expected response Y at all points of Y
  ➢ Implicitly assume that Y is a continuous variable

• Items are allowed to differ from each other in how much they relate to the latent trait, *but a good item is equally good for everybody*

• **CFA won’t work well for binary or categorical data**
  ➢ Thus, we need IRT
A History of “Common Factor Theory” (CFA)

• 1900’s :: Spearman’s G
  - Went looking for single-factor model... and “found” it
  - Led to development of other IQ tests (Stanford-Binet, Wechsler)

• 1930’s and 1940’s :: Thurstone elaborated Spearman’s model into a “multiple factor” model
  - Beginnings of exploratory factor analysis to do so
  - Later applied in other personality tests (e.g., MMPI)

• 1940’s and 1950’s: Guttman’s work
  - Factor analysis and test development is about generalizing from measures we have created to more measures of the same kind
  - Thus, need to think about structure before-hand
Common Factor Theory, continued

• 1940’s: Lawley provided a rigorous foundation for statistical treatment of common factor analysis
  ➢ But had to wait for better computers to be able to implement methods

• 1952: Lawley provided the beginnings of the confirmatory factor model
  ➢ Later extended by Howe and Bargmann (1950’s)
  ➢ Further extended by Jöreskog (LISREL – 1970’s)

• But this linear model should not be applied to dichotomous (or categorical) responses...
  ➢ Probability of correct response will go out of bounds
  ➢ Errors can’t be normally distributed with constant variance

• Enter Item Response Theory
  ➢ IRT is CFA for categorical variables
  ➢ The field of IRT is an example of generalized models
Item Response Theory (IRT)

- IRT resulted from combination of ideas from factor analysis and phi-gamma law of psychophysics
  - When detecting stimuli of varying intensity (e.g., light), the response follows a smooth, S-shaped curve that can be represented by the cumulative normal distribution
  - That response function also works to model probability of a correct response given (1 to 4) model parameters

- 1950: Lazarsfeld introduced “latent structure analysis”
  - Essentially a form of factor analysis for dichotomous items
  - Formed the beginnings of item response theory

- 1952: Lord introduced two-parameter normal ogive model
  - Now called an item factor model
  - Precursor to more common models today
Revisiting Our Regression Review

• Consider the following scenario:
  - You wish to predict Y from X, **BUT**
    - Y is now binary (can be either 0 or 1)
    - X is still continuous

• In this case, traditional regression will not work
Binary versus Continuous Outcome Variables

- Variable types:
  - Continuous: ranges from negative infinity to infinity
  - Binary: 0/1

- Means:
  - Continuous outcome mean: $\bar{Y}$
  - Binary outcome mean: proportion of 1’s = $p_Y$

- Variances:
  - Continuous: $\text{Var}(Y) = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1}$
  - Binary: $\text{Var}(Y) = p_Y(1 - p_Y) = p_Y q_Y = s_Y^2$

  * The variance IS determined by the mean!

<table>
<thead>
<tr>
<th>TABLE 3.2</th>
<th>Binary Item Variance and Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>p variance</td>
<td>.09</td>
</tr>
</tbody>
</table>
A Linear Model for Binary Outcomes

• If your outcome variable is binary (0 or 1):
  ➢ Expected mean is proportion of people who have a 1 (or “p”, the probability of Y=1)
    • The probability of having a 1 is what we’re trying to predict for each person, given the values on the predictors

• Under the regression model: $Y = \beta_0 + \beta_1 X + e$
  ➢ $\beta_0 =$ expected probability when all predictors are 0
  ➢ $\beta_1 =$ expected change in probability for a one-unit change in the predictor
  ➢ $e =$ difference between observed and predicted values

• Model becomes $Y = (\text{predicted probability of 1}) + e$
A Linear Model for Binary Outcomes

• But if Y is binary, then e can only be two things:
  - $e = \text{Observed } Y \text{ minus Predicted } Y$
    • If $Y = 0$ then $e = (0 - \text{predicted probability})$
    • If $Y = 1$ then $e = (1 - \text{predicted probability})$

• Mean of errors would still be 0...

• Variance of errors cannot be constant over levels of $X$
  like we assume in general linear models
  - The mean and variance of a binary outcome are dependent
  - This means that because the conditional mean of $Y$
    ($p$, the predicted probability $Y= 1$) is dependent on $X$,
    \textit{then so is the error variance}
A Linear Model for Binary Outcomes

- Needed: a method to translate probabilities bounded by zero and one to the entire number line
- Options:
  - Ignore bounding and use traditional general linear model
  - Transform probability to something continuous
3 Problems with Linear Regression Models for Binary Outcomes

1. Restricted range (e.g., 0 to 1 for binary item)
   - Predicted values can each only be off in two ways
     - So residuals can’t be normally distributed

2. Variance is dependent on the mean, and not estimated
   - Fixed and random parts are related
     - So residuals can’t have constant variance

3. Residuals have a limited number of possible values
   - Predicted values can each only be off in two ways
     - So residuals can’t be normally distributed
Differing Types of Outcomes

- **Generalized Linear Models** are General Linear Models
  - with differently distributed error terms
  - with transformed outcome variables

- Many kinds of non-normally distributed outcomes have some kind of generalized linear model to go with them:
  - **Binary (dichotomous)**
  - Unordered categorical (nominal)
  - Ordered categorical (ordinal)
  - Counts (discrete, positive values)
  - Censored (piled up and cut off at one end – left or right)
  - Zero-inflated (pile of 0’s, then some distribution after)

These two are often called “multinomial” inconsistently.
Parts of a Generalized Linear Model

- **Link Function (main difference from GLM):**
  - How a non-normal outcome gets transformed into something that is continuous (unbounded)
  - For outcomes that are already normal, general linear models are just a special case with an “identity” link function (Y * 1)
Generalized Models for Binary Outcomes

• Rather than modeling the probability of a 1 directly, we need to transform it into a more continuous variable with a link function, for example:

  ➢ Transform probability into an odds ratio:
    • Odds ratio: \( (p / 1-p) = \text{prob}(1) / \text{prob}(0) \)
    • If \( p = .7 \), then Odds(1) = 2.33; Odds(0) = .429
    • Odds scale is way skewed, asymmetric, and ranges from 0 to infinity

  ➢ Take natural log of odds ratio: called “logit” link
    • LN \( (p / 1-p) \): Natural log of \( (\text{prob}(1) / \text{prob}(0)) \)
    • If \( p = .7 \), then LN(Odds(1)) = .846; LN(Odds(0)) = - .846
    • Logit scale is now symmetric about 0
Model Background

• The log-odds is called a logit

\[
\text{Logit}(P(Y = 1)) = \ln \left( \frac{P(Y = 1)}{1 - P(Y = 1)} \right)
\]

• The logit is used because the responses are binary
  • Responses are either (1) or (0)
More on Logits

<table>
<thead>
<tr>
<th>Probability</th>
<th>Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>0.9</td>
<td>2.2</td>
</tr>
<tr>
<td>0.1</td>
<td>-2.2</td>
</tr>
<tr>
<td>0.99</td>
<td>4.6</td>
</tr>
</tbody>
</table>
From Logits to Probabilities

- Whereas logits are useful as the are unbounded continuous variables, categorical data analyses rely on estimated probabilities.

- The inverse logit function converts the unbounded logit to a probability.
  - This is also the form of an IRT model (and logistic regression).

\[ P(Y = 1) = \frac{\exp(\text{Logit}(P(Y = 1)))}{1 + \exp(\text{Logit}(P(Y = 1)))} \]
Non-Linearity in Prediction

- The relationship between X and the $P(Y = 1)$ is “non-linear”
  - An s-shaped logistic curve whose shape and location are dictated by the estimated model parameters (slope, intercept)
  - Linear with respect to the logit, non-linear with respect to probability

- The logit version of the model will be easier to explain; the probability version of the prediction will be easier to show
The Logistic Model

• Outcome is log odds (logit) of probability instead of probability
  ➢ Symmetric, unbounded outcome
  ➢ Assume linear relationship between predictors and log odds (logit)
  ➢ This allows an overall non-linear (S-shaped) relationship between X’s and probability of Y=1

• Errors are not assumed to be normal with constant variance
  ➢ ‘e_i’ will be missing – residual variance is NOT estimated
  ➢ Errors are assumed to follow a logistic distribution with a known residual variance of \( \pi^2/3 \) (3.29)
  ➢ Still assume errors are independent
    • Clustered data would need a generalized mixed model that would include random effects that account for any dependency
    • Items are like clustered data – items are typically treated as being nested within a person
Item Response Theory (IRT) Models

- Linear regression is to confirmatory factor models as to:
  - Logistic regression is to binary IRT models
  - Ordinal/nominal regression is to polytomous IRT models
  - IRT = Regression model predicting each categorical observed outcome variable from a latent variable(s) by using link functions

- “Rasch models” are a subset of IRT models with more restrictive assumptions...(but don’t let Rasch people hear you saying that)
  - The cult of Rasch: [http://www.rasch.org](http://www.rasch.org)

- Uses of IRT models:
  - *Correctly* analyze item-level data (binary items, Likert scales)
  - Examine sensitivity of measurement across range of latent trait
  - Provide comparability across persons, items, and occasions
Example Item Response Curves

\[ P(u = 1 | \theta) \]

- **a = Discrimination**: slope of ‘line’
- **b = Difficulty**: location of ‘line’
- **c = Lower Asymptote of ‘line’**
- **d = Upper Asymptote of ‘line’**

PRE 932: Lecture #1
Item Response Theory, continued

• IRT unit of analysis is the individual ITEM
  ➢ Nonlinear response model that simultaneously accounts for differences between persons AND differences between items
    • Items and persons are put on the same latent metric
    • Probability of getting an item right depends (at least) on the subject’s ability and the item’s difficulty
    • Ability is interpreted relative to item performance, not (just) relative to other people in the sample
  ➢ All items are NOT created equal (not exchangeable)
    • Having items that differ in their properties is a GOOD THING
  ➢ Error is not a static characteristic of the test
    • Reliability varies across ability level, and depends specifically on how well the items match the subjects (i.e., on information)
Item Response Theory, continued

- 1952: Lord’s seminal paper: Spearman’s single-factor model can be applied to dichotomous items
  - Dichotomous responses modeled by normal ogive function
  - Elaborated in 1960’s by Birnbaum :: Transform outcome using logit link, assume Bernoulli error

- 1968: Lord & Novick → first CTT text to also include IRT
  - Well-connected to emerging scholars in both educational testing and psychometric methods

- 1960: Separate line of development by Rasch (no ‘a’/factor loading parameter)
  - Restricted IRT model, but with highly desirable properties
  - ... and somewhat different philosophical viewpoint
Unified View of Test Theory
(courtesy of McDonald, 1999)

- Classical test theory can be viewed as a restricted form of the common factor model, but the focus is the TEST...
  - Originated by Spearman, elaborated by Thurstone, formalized by Lawley, and made practical by Jöreskog

- Item response (and Rasch) models for dichotomous data are basically nonlinear common factor models...
  - Developed by Lord, Birnbaum, and Rasch and their students

- Common factor models (CFA) are a linear approximation to the item response model when applied to dichotomous or ordinal responses
  - Approximation with varying degrees of success

- Other newer measurement models to measure latent traits
  - Count, zero-inflated, two-part....
Advantages of the Measurement Model Framework (CFA, IRT, and beyond)

• Explicit, testable models of dimensionality
• Concrete guidelines for selecting items to build scales
• Assess measurement sensitivity across range of latent trait (i.e., know where the ‘holes’ are)
• Provide comparability across persons, items (different forms scales or different scales), and occasions
• Examine comparability across distinct groups (perhaps bias exists)
  ➢ Confirmatory factor analysis :: “Measurement invariance”
  ➢ Item response theory :: “Differential item functioning”
• Internal and external evidence for construct validity
• Flexible measurement models for different response formats and distributions (CFA, IRT, and others)
Disadvantages of Framework (CFA and IRT)

• Primary: Required sample size
  ➢ Casts of 100s for sure, and preferably 1000s
  ➢ Uses maximum likelihood (although WLSMV in Mplus can now be used for multidimensional IRT models on fewer cases)

• Technical difficulties
  ➢ Estimation difficulties
  ➢ Fighting with software
  ➢ References written in Greek (literally)
Summary: Psychometric Introduction

- Test Theory is a collection of statistical models used to evaluate the quality of an instrument in measuring a latent trait
  - “Classical Test Theory” (CTT)
    - Just add items up: Focus on TEST as unit of analysis
    - Simple, yet very restrictive; requires belief instead of evidence
  - “Latent Trait Models” (CFA, IRT... and beyond)
    - Estimate a latent trait; Focus on ITEM as unit of analysis
    - Flexible models that differ by response format of items
    - More complex, but more powerful and useful

- The nuances of IRT are due to the nature of modeling categorical data and the needs of the fields that are using the methods
CONCLUDING REMARKS
IRT :: CFA as Logistic Regression :: Linear Regression

<table>
<thead>
<tr>
<th>Outcome Type</th>
<th>Observed X</th>
<th>Latent X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous Y</td>
<td>Linear Regression</td>
<td>Confirmatory Factor Models</td>
</tr>
<tr>
<td>“General Linear Model”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete Y</td>
<td>Logistic Regression</td>
<td>Item Response Models</td>
</tr>
<tr>
<td>“Generalized Linear Model”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The basis of Item Response Theory lies in models for discrete outcomes, which are called “generalized” models.
- Thus, IRT and CFA seek to achieve the same results with different types of data.
# How Diagnostic Models Fit

<table>
<thead>
<tr>
<th>Outcome Type</th>
<th>Latent X: Continuous</th>
<th>Latent X: Ordered Categorical</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Continuous Y</strong></td>
<td>Confirmatory Factor Models</td>
<td>Our simulation-based study (hint: Diagnostic Classification Models)</td>
</tr>
<tr>
<td>“General Linear Model”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Discrete Y</strong></td>
<td>Item Response Models</td>
<td>Diagnostic Classification Models</td>
</tr>
<tr>
<td>“Generalized Linear Model”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>