Comparing IRT with Other Models

Lecture #14

ICPSR Item Response Theory Workshop

Lecture Overview

- The final set of slides will describe a parallel between IRT and another commonly used method for measurement: factor analysis

- These slides are meant to provide a basis for comparing the two methods, including the appropriate times for applying each method

MORE COMPARING IRT WITH CFA

Introduction

- Consider a score on an item that is not categorical
  - Rather, consider a score to be continuous
  - For simplicity, call the item X_i
- Often Likert-type item scores are considered continuous
  - Other examples of continuous item types include reaction time and many physiological measurements
- Our goal will be to model the response behavior of an examinee on item X_j using a latent variable (in IRT, typically called θ)
  - To distinguish the two approaches, we will use F as the latent variable for factor analysis

The Spearman Single Factor Model

- In relating an examinee's level of the common factor to their performance on an item, we introduce the Spearman single factor model:

$$X_{is} = \mu_i + \lambda_i F_s + E_{is}$$

- X_{is} is the score for a person s on the i^{th} item
- F_s is the persons's factor score

The Spearman Single Factor Model

$$X_{is} = \mu_i + \lambda_i F_s + E_{is}$$

- E_{is} is the unique or idiosyncratic property of item
  - The amount by which item *i* is shifted from the predicted value for person *e*
- μ_i is the overall mean for an item
  - Allowing for differing item difficulties
- λ_i is called the factor loading of item *i*
  - We will discuss this factor loading quite a bit

Factor Loadings

- The term factor loading has a long history in Psychology
- The extent to which the item is “loaded” onto the factor
  - Some items load more highly on to the factor than other
- The factor loadings of items reveal much about a test’s structure :: the mapping of items onto factors

More on Factor Loadings

- The factor loading is similar to a regression weight:
  - It represents the amount of change in the item per one-unit increase in the factor score
- It measures how sensitively each item functions as an indicator of the common factor F
  - Items with relatively large loadings are better indicators of F than items with relatively small loadings
- The factor loading is a measure of the discriminating power of the item
  - How well the item discriminates between examinees with low and high values of F

Single Factor Model Specifics

- We need to define a few more things about our factor model:
  - The unique component, E_{is}, is independent of the common factor, F_s
    - Independence means that Cov(E,F) = Corr(E,F) = 0
  - The unique components of any two items i and j are independent:
    - $\text{Cov}(E_{is}, E_{js}) = \text{Corr}(E_{is}, E_{js}) = 0$
  - The mean for the unique component is zero

More Specifics

- Like in IRT, we also have to set the scale for F
  - We must pick it's mean and variance
  - For most of our purposes, it serves us well to think of F as being a standardized measure
    - Mean of zero
    - Standard Deviation/Variance of one

- Standardized factors aren't as common in factor analysis (CFA or SEM)
  - Variance of the factor must be fixed when additional modeling features are added
    - Actually, same in IRT

What Does The Common Factor Model Say About Our Items?

- So, what can we say the model predicts about our items, marginally?
- What is the model-predicted item mean?
- What is the model-predicted item variance?
- Why are these important?

Model Predicted Item Mean

- The mean for an item under the single factor model:

$$
\begin{aligned}
E(X_{is}) &= E(\mu_i + \lambda_i F_s + E_{is}) \\
&= E(\mu_i) + E(\lambda_i F_s) + E(E_{is}) \\
&= \mu_j + \lambda_i E(F_s) + E(E_{is}) \\
&= \mu_i + \lambda_i * 0 + 0 \\
&= \mu_i
\end{aligned}
$$

Item Mean is Trivial

- The factor model says that our item mean should be our item mean parameter

- Generally, we are not concerned with such a quantity because it tells us information only marginally
  - No information about how the item measures the common factor

Model Predicted Item Variance

- The variance for an item under the single factor model:

$$\text{Var}(X_{is}) = \text{Var}(\mu_i + \lambda_i F_s + E_{is})$$

$$= \text{Var}(\lambda_i F_s + E_{is})$$

$$= \text{Var}(\lambda_i F_s) + \text{Var}(E_{is}) + 2\,\text{Cov}(F_s, E_{is})$$

$$= \lambda_i^2\,\text{Var}(F_s) + \text{Var}(E_{is})$$

$$= \lambda_i^2 + \psi_i^2$$

Is zero by independence

We Typically Set this to One

We define the variance of E to be the unique variance of the item.

Model Predicted Item Variance

- The variance for an item under the single factor model:

$$\begin{aligned}
\text{Var}(X_{is}) &= \text{Var}(\mu_i + \lambda_i F_s + E_{is}) \\
&= \text{Var}(\lambda_i F_s + E_{is}) \\
&= \text{Var}(\lambda_i F_s) + \text{Var}(E_{is}) + 2\,\text{Cov}(F_s, E_{is}) \\
&= \lambda_i^2\,\text{Var}(F_s) + \text{Var}(E_{is}) \\
&= \lambda_i^2 + \psi_i^2
\end{aligned}$$

→ We define the variance of E to be the unique variance of the item.

Model Predicted Item Covariances

- The covariance for a pair of items under the single factor model :

$$\text{Cov}(X_i, X_j) = \text{Cov}(\mu_i + \lambda_i F_s + E_{is}, \mu_j + \lambda_j F_s + E_{js})$$
$$= \text{Cov}(\lambda_i F_{is} + E_{is}, \lambda_j F_s + E_{js})$$
$$= \text{Cov}(\lambda_i F_s, \lambda_j F_s) + \text{Cov}(\lambda_i F_s, E_{js}) + \text{Cov}(\lambda_j F_s, E_{is}) + \text{Cov}(E_{is}, E_{js})$$
$$= \lambda_i \lambda_j \text{Cov}(F_s, F_s)$$
$$= \lambda_i \lambda_j$$

The covariance of a variable with itself is its variance.

The variance of F is set to one.

Extrapolating to the Covariance Matrix

- We have seen:
  - The model predicted variance for each item
  - The model predicted covariance for each pair of items

- The model-predicted covariance matrix looks like:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \lambda_1^2 + \psi_1^2 & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 & \lambda_1\lambda_5 \\ \lambda_1\lambda_2 & \lambda_2^2 + \psi_2^2 & \lambda_2\lambda_3 & \lambda_2\lambda_4 & \lambda_2\lambda_5 \\ \lambda_1\lambda_3 & \lambda_2\lambda_3 & \lambda_3^2 + \psi_3^2 & \lambda_3\lambda_4 & \lambda_3\lambda_5 \\ \lambda_1\lambda_4 & \lambda_2\lambda_4 & \lambda_3\lambda_4 & \lambda_4^2 + \psi_4^2 & \lambda_4\lambda_5 \\ \lambda_1\lambda_5 & \lambda_2\lambda_5 & \lambda_3\lambda_5 & \lambda_4\lambda_5 & \lambda_5^2 + \psi_5^2 \end{bmatrix}$$

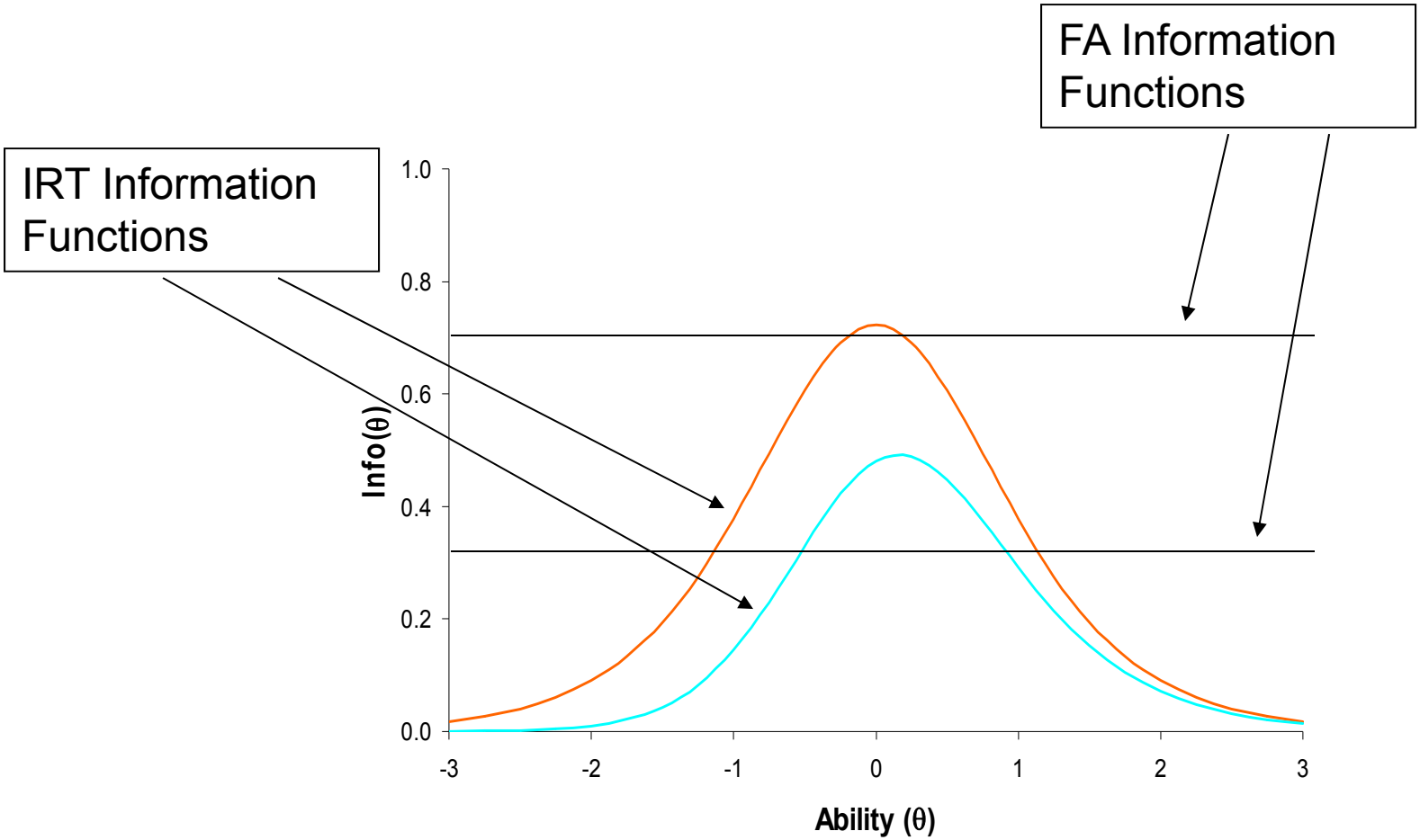
Item Information Under the Factor Model

- The item information function under the single factor model is:

$$I(X_i) = \frac{\lambda_i^2}{\psi_i^2}$$

- Item information under the factor model is not a function of the latent trait
  - This is a key distinction between item information in the factor model and IRT
  - It is a consequence of the differences in data type
    - A nuance of categorical data: the mean and the variance are related

Item Information Functions



Implications of FA Item Information

- The test information function in FA is flat
  - Regardless of a person's location on the scale, the standard error of their estimate will be the same

- CAT algorithms do not make much sense using FA
  - All items would be equally informative across the scale
  - The items with the highest information would always be selected

Comparing FA with IRT

- FA and IRT have much in common:
  - They both provide a statistical model for response behavior as a function of a latent trait (or set of latent traits)

- IRT parameterizations obscure the commonalities between the models
  - To demonstrate, let's rephrase the 2PL model

IRT Model in Slope/Intercept Form

- Begin with the original 2PL Model:

$$P(Y_{is} = 1|\theta_s) = \frac{\exp\left(1.7a_i(\theta_s - b_i)\right)}{1 + \exp\left(1.7a_i(\theta_s - b_i)\right)}$$

- Then convert into the log-odds of the probability of a correct response:

$$\ln\left(\frac{P(Y_{is} = 1|\theta_s)}{1 - P(Y_{is} = 1|\theta_s)}\right) = 1.7a_i(\theta_s - b_i)$$

IRT Model in Slope/Intercept Form

- Finally, multiply through the equation:

$$\ln\left(\frac{P(Y_{is}=1|\theta_s)}{1-P(Y_{is}=1|\theta_s)}\right)=1.7a_i(\theta_s-b_i)=-1.7a_ib_i+1.7a_i\theta_s$$

- Now, we can re-configure terms to FA analogs:

$$\ln\left(\frac{P(Y_{is}=1|\theta_s)}{1-P(Y_{is}=1|\theta_s)}\right)=1.7a_i(\theta_s-b_i)=-1.7a_ib_i+1.7a_i\theta_s$$
$$=\mu_i+\lambda_i\theta_s$$

IRT vs. FA

- Many IRT models are categorical versions of the FA or structural equation model
  - The difference in model properties (like information) is due to the link function used for the data

- A link function is the function applied to the left hand side of the previous equation
  - IRT models we have discussed usually use a logistic link function (or an ogive)
  - FA models use an "identity" link function
    - Identity = no link function at all

IRT vs. FA, continued

- Appropriate uses of FA are for data that follow continuous distributions

- Appropriate uses of IRT are for data that follow the corresponding categorical distribution
  - Binary variables use binomial logistic
  - Polytomous variables use multinomial logistic

- The question to be asked is at what point do categorical data become continuous
  - If you think really hard about it, all data are categorical…how many categories, though?

Other Link Functions: Item Factor Analysis

- Just as in categorical data analysis, other link functions exist and their use results in models with IRT-like properties

- One of the more prevalent link functions is the probit or normal ogive link
  - This is the cumulative distribution function of a standard normal variable
  - The use of the normal ogive link dates to Lord (1952)
  - More commonly, such models are referred to as Item Factor Models
  - Setting the scaling constant to 1.7 in IRT approximates this function

Item Factor Analysis

- An alternative parameterization of the model in terms of underlying quantitative "response tendencies" – common factor parameterization

- Each binary item has associated with it an "underlying" quantitative response tendency X_i^* and a threshold value τ_i , such that:
  - If $X_i^* > \tau_i$ then $X_i = 1$
  - If $X_i^* \leq \tau_i$ then $X_i = 0$

Underlying Response Model

- The underlying response tendencies, $X_1^*, \ldots, X_I^*$ are then used with a factor analysis model, say Spearman single factor model (mean omitted – see below):

$$X_i^* = \lambda_i F_s + E_{is}^*$$

- With uncorrelated unique parts E_{is}^*

- For model identification, we impose a scale on each X_i^* so that it is standardized:
  - With mean zero (hence no μ_i).
  - With variance one, so $\lambda_i^2 + \psi_i^2 = 1$.

Building on the Previous Model

- F_s and E_{is}^* each have a normal distribution
- Each X_i^* has a normal distribution
- This leads to, for an item i,

$$P(X_{is} = 1|F_s) = P(X_{is}^* > \tau_i|F_s)$$
$$= \Phi\left(\frac{\lambda_i}{\sqrt{1-\lambda_i^2}}F_s - \frac{1}{\sqrt{1-\lambda_i^2}}\pi_i\right)$$

- Larger λ_i means larger discriminating power
- The larger the π_i, the more difficult the item

Similarities

- We can relate our new item factor analysis parameterization to the IRT parameterization:

	Item Factor Analysis	IRT
Item Discrimination	$\lambda_i = \frac{b_i}{\sqrt{1 + b_i^2}}$	$a_i = \frac{\lambda_i}{\sqrt{1 - \lambda_i^2}}$
Item Difficulty	$\pi_i = -\frac{a_i}{\sqrt{1 + b_i^2}}$	$b_i = -\frac{\pi_i}{\sqrt{1 - \lambda_i^2}}$

So Why Use One Parameterization over the Other?

- The common factor parameters are most useful in a preliminary examination of the structure of the data
  - Many people are experienced with using factor loadings
  - Because we can use established factor-analytic criteria for judging the sizes of the factor loadings.

- The response function parameterizations are useful in applications of a fitted model because they generally simplify computations
  - Differing estimation routines can be employed

Interpretation of Item Factor Parameters

- One can interpret the common factor parameters in relation to classical item analysis:

- The factor loading of the item, λ_i, is the product-moment correlation between X_i^* and F
  - Which is the biserial correlation between binary X_i and F

- The product of the factor loadings between any pair of items (i and j) gives the model estimate of the tetrachoric correlation between the items:

$$\rho_{ij} = \lambda_i \lambda_j$$

GENERALIZING BEYOND CATEGORICAL AND CONTINUOUS DATA

Welcome to the Family

- **General*ized* Linear Models** → General Linear Models with non-normal error terms and transformed data to obtain some kind of continuous outcome to work with

- Many kinds of non-normally distributed outcomes have some kind of generalized linear model to go with them:
  - Binary (dichotomous)
  - Ordered categorical (ordinal)
  - Unordered categorical (nominal)

  } These two are often called "multinomial" inconsistently

  - Censored (piled up and cut off at one end – left or right)
  - Counts (discrete, positive values)
  - Counts with zero issues (too many or none)

3 Parts of a Generalized Linear Model

- **Link Function (main difference from GLM):**
  - How a non-normal outcome gets transformed into something we can predict that is more continuous (unbounded)
  - For outcomes that are already normal, general linear models are just a special case with an "identity" link function (Y * 1)

- **Model for the Means ("Structural Model"):**
  - How predictors linearly relate to the transformed outcome
  - New transformed data = $\beta_0 + \beta_1 X_s + \beta_2 Z_s$

- **Model for the Variance ("Sampling/Stochastic Model"):**
  - If the errors aren't normal and homoscedastic, then what are they?
  - Family of alternative distributions at our disposal that map onto what the distribution of errors could possibly look like

Model Parts for Binary Outcomes: 2 Choices → Logit vs. Probit

- 2 Alternative Link Functions:
  - **Logit link: binary Y = ln(p/1-p) :: logit is new transformed Y**
    - Y is 0/1, but logit(Y) goes from $-\infty$ to $+\infty$
  - **Probit link: binary Y = Φ(Y)**
    - Observed probability replaced by value of standard normal curve below which observed proportion is found :: **Z-score is new transformed Y**
    - Y is 0/1, but probit(Y) goes from $-\infty$ to $+\infty$

- Same Model for the Means:
  - Main effects and interactions of predictors as desired…
  - No analog to odds coefficients in probit, however

- 2 Alternative Models for the Variances:
  - **Logit:** e_s's ~ Bernoulli distributed with known variance of $\boldsymbol{\pi^2/3}$**, or 3.29**
  - **Probit:** e_s's ~ Bernoulli distributed with known variance of **1**

Ordered Categorical Outcomes

- One option: **Cumulative Logit Model**
  - Called "graded response model" in IRT
  - Assumes ordinal categories
  - Model logit of category vs. all lower/higher via submodels
    - 3 categories :: 2 models: 0 vs. 1 or 2, 0 or 1 vs. 2
  - Get separate threshold (-intercept) for each submodel
  - Effects of predictors are assumed the same across submodels :: "Proportional odds assumption"
    - Is testable in some software (e.g., Mplus, NLMIXED)
  - In Mplus, can do this with the **CATEGORICAL ARE** option

Ordered Categorical Outcomes

- Another option: **Adjacent Category Logit Model**
  - Called "partial credit model" in IRT
  - Does not assume order across all categories (only adjacent)
  - Model logit of sequential categories only via submodels
    - 3 categories :: 2 models: 0 vs. 1, 1 vs. 2
  - Get separate threshold (-intercept) for each submodel
  - Effects of predictors are still assumed the same across adjacent category submodels :: "Proportional odds assumption"
    - Is testable in some software (e.g., NLMIXED)
  - Currently not available in Mplus

Unordered Categorical Outcomes: "Nominal Model"

- Compare each category against a reference category using a binary logit model
  - Referred to as "baseline category logit"

- End up with multiple logistic submodels up to #categories – 1
  (2 submodels for 3 categories, 3 for 4 categories, etc)

- Intercept/thresholds and slopes for effects of predictors (factor loadings) are estimated separately within each binary submodel
  - Can get effects for missing contrast via subtraction
  - Effects are interpreted as "given that it's one of these two categories, which has the higher probability"?

- Model comparisons proceed as in logistic regression
  - Can also test whether outcome categories can be collapsed

- In Mplus, can do this with the **NOMINAL ARE** option

Censored ("Tobit") Outcomes

- For outcomes with ceiling or floor effects
  - Can be "Right censored" and/or "left censored"
  - Also "inflated" or not ::
    - inflation = binary variable in which 1 = censored, 0 = not censored

- Model assumes unobserved continuous distribution instead for the part it is missing

- In Mplus, can do with various CENSORED ARE (with options):
  - CENSORED ARE y1 (a) y2 (b) y3 (ai) y4 (bi);
    - y1 is censored from above (right); y2 is censored from below (left)
    - y3 is censored from above (right) and has inflation variable (inflated: y3#1)
    - y4 is censored from above (below) and has inflation variable (inflated: y4#1)
  - So, can predict distribution of y1-y4, as well as whether or not y3 and y4 are censored ("inflation") as separate outcomes
    - y3 ON x; → x predicts value of Y if at censoring point or above
    - y3#1 ON x; → x predicts whether Y is censored (1) or not (0)

A Family of Options in Mplus for Count Outcomes (COUNT ARE)

- Counts: non-negative integer unbounded responses
  - e.g., how many cigarettes did you smoke this week?

- Poisson and negative binomial models
  - **Same Link: count Y = ln(Y)** (makes the count stay positive)
  - **$LN(Y_{is}) = \mu_i + \lambda_i F_s + e_{is}$** (model has intercepts and loadings)
  - Residuals follow 1 of 2 distributions:
    - Poisson distribution in which k = Mean = Variance
    - Negative binomial distribution that includes a new α "scaling" or "over-dispersion" parameter that allows the variance to be bigger than the mean :: variance = k(1 + kα)
    - Poisson is nested within negative binomial (can test of $\alpha \neq 0$)
    - COUNT ARE y1 (p) y2 (nb); :: y1 is Poisson; y2 is neg. binomial

Issues with Zeros in Count Data

- No zeros :: zero-truncated negative binomial
  - e.g., how many days were you in the hospital? (has to be >0)
  - COUNT ARE y1 (nbt);

- Too many zeros :: zero-inflated poisson or neg binomial
  - e.g., # cigarettes smoked when asked in non-smokers too
  - COUNT ARE y2 (pi) y3 (nbi);
    - Refer to "inflation" variable as y2#1 or y3#1
  - Tries to distinguish 2 kinds of zeros
    - "Structural zeros" – would never do it
      - Inflation is predicted as logit of being a structural zero
    - "Expected zeros" – could do it, just didn't (part of regular count)
      - Count with expected zeros predicted by poisson or neg binomial
  - Poisson or neg binomial without inflation is nested within models with inflation (and poisson is nested within neg binomial)

Issues with Zeros in Count Data

- Other more direct ways of dealing with too many zeros: split distribution into (0 or not) and (if not 0, how much)?
  - Negative binomial "hurdle" (or "zero-altered" negative binomial)
    - COUNT ARE y1 (nbh);
    - 0 or not: predicted by logit of being a 0 ("0" is the higher category)
    - How much: predicted by zero-truncated negative binomial
  - Two-part model uses Mplus DATA TWOPART: command
    - NAMES ARE y1-y4; → list outcomes to be split into 2 parts
    - CUTPOINT IS 0; → where to split observed outcomes
    - BINARY ARE b1-b4; → create names for "0 or not" part
    - CONTINUOUS ARE c1-c4; → create names for "how much" part
    - TRANSFORM IS LOG; → transformation of continuous part
    - 0 or not: predicted by logit of being NOT 0 ("something" is the 1)
    - How much: predicted by transformed normal distribution (like log)

CONCLUDING REMARKS

Wrapping Up…

- When fitting latent factor models (or when just predicting observed outcomes from observed predictors instead), you have many options to fit non-normal distributions
  - **CFA:** Continuous outcomes with normal residuals, X → Y is linear
  - **IRT and IFA:** Categorical or ordinal outcomes with Bernoulli/multinomial residuals, X → transformed Y is linear; X → original Y is nonlinear
  - **Censored:** Continuous outcomes that shut off, X → Y is linear
    - Model tries to predict what would happen if Y kept going instead
  - **Count family:** Non-negative integer outcomes, X → LN(Y) is linear
    - Residuals can be Poisson (where mean = variance) or negative binomial (where variance > mean); either can be zero-inflated or zero-truncated
    - Hurdle or two-part may be more direct way to predict/interpret excess zeros (predict zero or not and how much rather than two kinds of zeros)