The use of test scores in secondary analysis

Irini Moustaki
London School of Economics and Political Science

Framework and notation

- The latent variables used as predictors or outcomes could be either continuous (IRT, factor analysis models) or categorical (latent class models)
- The regression could be linear or non-linear.
- Here the focus is on distal outcome regressions.
- Notation:
  - X is a set of indicators that measure the latent variables θ.
  - Y is a distal outcome
  - Z is a set of covariates
- Bias in the regression estimates from the model that involves factor scores as predictors or outcomes.
- The bias is determined by the amount of measurement error in the estimated θ.

The proposed model - one step approach

$$f(Y, X \mid Z) = \int f(Y \mid X, \theta, Z) f(X \mid \theta, Z) f(\theta \mid Z) d\theta$$

- Measurement equivalence: $f(X \mid \theta)$
- Distal outcome regression (structural model): $f(Y \mid \theta, Z)$
- Structural model/ conditioning model: $f(\theta \mid Z)$

$$f(Y, X \mid Z) = \int f(Y \mid \theta, Z) f(X \mid \theta) f(\theta \mid Z) d\theta$$

- This is what the author call the Mixed Effects Structural Equations (MESE) model.
- MESE is found to be robust under misspecification of the IRT model.

Modelling approaches

- One-step approach and various stepwise approaches (including the modal class, modied Bolck,Croon, Hagenaars [BCH], Lanza, Tan, Bray [LTB], and three-step maximum likelihood [ML] methods).
- One-step approach is the one described in the presentation (MESE).
  - The One-step approach simultaneously estimates the measurement model and the regression model of Y on θ, treating Y as an additional indicator for θ.
  - Parameter estimates, and regression coefficients, are obtained by jointly maximizing the log-likelihood of response patterns and the distal outcome.
  - The Y variable is included in the measurement model together with the X variables.

Two-step approach

- Step 1: the measurement model is estimated alone
- Step 2: the parameters of this measurement model are held fixed when the structural model is estimated.
- Estimated standard errors are derived for the two-step estimates of the structural model which account for the uncertainty from both steps of the estimation.

Reference: Two-Step Estimation of Models Between Latent Classes and External Variables (Bakk and Kuha, 2018).

Three-step approach

- Step 1: Perform a latent class analysis without Y or Z. Calculate the posterior probability of being in each class and the modal class M for each individual.
- Step 2: Calculate the misclassification probabilities which will be treated as fixed quantities in Step 3.
- Step 3: Maximize the log-likelihood function that weights the observed data by the mis-classification probabilities.

The three-step approach has been recently advocated over the simultaneous one-step approach to model a distal outcome predicted by a latent categorical variable.

One-Step Approach - Advantages

- It is more efficient compared to stepwise approaches that might introduce additional uncertainty between steps.
- It allows for more flexible model structures, such as models with direct effects of covariates on indicators and the distal outcome.
- It is straightforward to account for residual correlation between Y and Xs, beyond that captured by class membership (Bakk et al., 2013).

One-Step Approach - Disadvantages

- We require Y to be conditionally independent from the other indicators (X).
- Vermunt (2010) noted the burden of having to re-estimate the entire model should one decide to add or delete covariates in the measurement model.
- A more serious issue is the inclusion of a distal outcome into the measurement model creates an unintended circular relationship in that the latent variable θ that is supposed to explain Y is also determined partly by Y. If there are multiple distal outcomes, the shift in the latent class proportions can be severe, especially when the classes outnumber the indicators or when class separation is poor (low entropy).
- By treating Y as an indicator for θ, we require for continuous Y that is normally distributed given θ.

Some literature results (1)

- Simulation studies have found that the performance of the one- and three-step approaches are similar in most situations (Bakk & Vermunt, 2016).
- From studies that considered latent class models with latent variables as predictors of a distal outcome: when all necessary model assumptions hold, the sample size is large, and class separation is good: all methods perform well with small bias, correct standard errors, and good coverage.

Some literature results (2)

- Under various degrees of violation of the normality and conditional independence assumption for the distal outcome and indicators, both approaches are subject to bias but the three-step approach is less sensitive.
- When there is local dependence between Y and the indicators for the latent variables, the one-step approach leads to greater bias than the three-step ML approach. This is mainly explained by a tendency to extract too many classes when there is residual correlation between Y and the Xs. The extraction of pseudo-classes is not necessarily wrong from a theoretical point of view, but one needs to question the validity of such extra classes, which might not be interpretable.

Recent extensions

- A three-step approach where the distal outcome is predicted by multiple and possibly associated latent categorical variables (Zhu, Steele and Moustaki, SEM 2017)
- Two-Step Estimation of Models Between Latent Classes and External Variables (Bakk and Kuha, 2018).

The Roy model with latent variables

- Introduce latent variables in the Roy model of self-selection of outcomes.
- Compute ATE and TTE given latent variables.
- Three very interesting applications.

Some general comments

- It is not clear how the counterfactuals are being measured.
- The model resembles the causal inference framework of potential outcomes but what about the assumptions required for computing ATE in observational studies? (assumption of sequential ignorability).
- How are the latent variables being identified in the examples?
- Mixtures of distributions have been assumed for the latent variables but have not been justified. There is no discussion on model fit or model selection.
- How do you compute the ATE and TTE in the presence of latent variables?