3-Step ML Auxiliary Variable Integration Using `MplusAutomation`

Adding Covariate and Distal Outcome Variables to Mixture Models

IMMERSE Project: Adam Garber

February, 09, 2023

The Institute of Mixture Modeling for Equity-Oriented Researchers, Scholars, and Educators (`IMMERSE`) is an IES funded training grant (R305B220021) to support education scholars in integrating mixture modeling into their research.

Visit our Website to learn more about the IMMERSE project.

Follow us on Twitter for updates on posted resources!

Visit our GitHub account to follow along with this tutorial & others.

What is included in this video tutorial?

This `R` tutorial automates the 3-step ML axiliary variable procedure using the `MplusAutomation` package (Hallquist & Wiley, 2018) to estimate models and extract relevant parameters. To learn more about auxiliary variable integration methods and why multi-step methods are necessary for producing un-biased estimates see Asparouhov & Muthén (2014).

The motivation for this tutorial is that conducting the 3-step manually is highly error prone as it requires pulling logit values estimated in the step-1 model and adding them in the model statement of the step-2 model (i.e., lots of copying & pasting). In contrast, this approach is fully replicable and provides clear documentation which translates to more reliable research. Also, it saves time!

How to reference this tutorial:

Garber, A. C. (2021). 3-Step ML Auxiliary Variable Integration Using MplusAutomation. Retrieved from psyarxiv.com/phtxa

---

Follow along! Link to `Github` repository:

https://github.com/immerse-ucsb/3step-ML-auto

Load packages

```
library(MplusAutomation) # Conduit between R & Mplus
library(glue)            # Pasting R code into strings
library(here)            # Location, location, location
library(tidyverse)       # Tidyness
```

Data Source: Civil Rights Data Collection (CRDC)

> The CRDC is a federally mandated school and district level data collection effort that occurs every other year. This public data is currently available for selected variables across 4 years (2011, 2013, 2015, 2017) and all US states. In the following tutorial six focal variables are utilized as indicators of the latent class model; three variables which report on harassment/bullying in schools based on disability, race, or sex, and three variables on full-time equivalent school staff employees (counselor, psychologist, law enforcement). For this example, we utilize a sample of schools from the state of Arizona reported in 2017.

Information about CRCD: https://www2.ed.gov/about/offices/list/ocr/data.html

Data access (`R`): https://github.com/UrbanInstitute/education-data-package-r

---

Read in CSV data file from the `data` subfolder

```
bully_data <- read_csv(here("data", "crdc_aux_data.csv"))
```

“Manual 3-Step” ML Auxiliary Variable Integration Method

Step 1 - Estimate the unconditional model with all covariate & distal outcome variables mentioned in the `auxiliary` statement.

NOTE: In this example, Mplus input and output files are directed to the sub-folder `3step_mplus`. Due to the fact that adding auxiliary variables is conducted after enumeration, generally other sub-folders will exist in the top-most `Rproject` folder such as `enum_mplus`, `data`, and `figures`.

```
m_step1  <- mplusObject(
  TITLE = "Step1  (MANUAL 3-STEP ML APPROACH)",
  VARIABLE =
   "categorical = X1 X2 X3 X5 X6;

    usevar = X1 X2 X3 X5 X6;

    classes = c(3);

    !!! All auxiliary variables to be considered in the final model should be listed here !!!
    auxiliary =
    COVAR1 DISTAL1 DISTAL2;",

  ANALYSIS =
   "estimator = mlr;
    type = mixture;
    starts = 500 100;",

  SAVEDATA =
   "!!! This saved dataset will contain class probabilities and modal assignment columns !!!
    File=3step_savedata_012020.dat;
    Save=cprob;
    Missflag= 999;",

  MODEL = "",
  OUTPUT = "",

  PLOT =
    "type = plot3;
    series = X1 X2 X3 X5 X6(*);",

  usevariables = colnames(example_data),
  rdata = example_data)

m_step1_fit <- mplusModeler(m_step1,
                 dataout=here("3step_mplus", "Step1_3step.dat"),
                 modelout=here("3step_mplus", "Step1_3step.inp") ,
                 check=TRUE, run = TRUE, hashfilename = FALSE)
```

Step 2 - Extract logits & saved data from the step 1 unconditional model.

Extract logits for the classification probabilities for the most likely latent class

```
logit_cprobs <- as.data.frame(m_step1_fit[["results"]]
                                         [["class_counts"]]
                                         [["logitProbs.mostLikely"]])
```

Extract saved data from the step 1 model `mplusObject` named "m_step1_fit"

```
savedata <- as.data.frame(m_step1_fit[["results"]]
                                     [["savedata"]])
```

Rename the column in savedata for "C" and change to "N"

```
colnames(savedata)[colnames(savedata)=="C"] <- "N"
```

Step 3 (part 1) - Estimate the unconditional model with logits from step 2.

This model is estimated to check that the class proportions are approximately the same as in step 1.

```
m_step2  <- mplusObject(
  TITLE = "Step2  (MANUAL 3-STEP ML APPROACH)",

  VARIABLE =
 "nominal=N;
  USEVAR = n;
  missing are all (999);
  classes = c(3); ",

  ANALYSIS =
 "estimator = mlr;
  type = mixture;
  starts = 0;",

  MODEL =
    glue(
 "%C#1%
  [n#1@{logit_cprobs[1,1]}];
  [n#2@{logit_cprobs[1,2]}];

  %C#2%
  [n#1@{logit_cprobs[2,1]}];
  [n#2@{logit_cprobs[2,2]}];

  %C#3%
  [n#1@{logit_cprobs[3,1]}];
  [n#2@{logit_cprobs[3,2]}];"),

  OUTPUT = "!tech11  tech14 res;",
  PLOT =
 "!type = plot3;
  !series = X1 X2 X3 X5 X6(*);",

  usevariables = colnames(savedata),
  rdata = savedata)

m_step2_fit <- mplusModeler(m_step2,
               dataout=here("3step_mplus", "Step2_3step.dat"),
               modelout=here("3step_mplus", "Step2_3step.inp"),
               check=TRUE, run = TRUE, hashfilename = FALSE)
```

Step 3 (part 2) - Add covariates & distal outcomes to the model.

Estimate the final SEM Model - Moderation Example

Specification details:

- This example contains two distal outcomes (`DISTAL1` & `DISTAL2`) and one binary covariate (`COVAR1`).
- Under each class-specific statement (e.g., `%C#1%`) the distal outcomes are mentioned to estimate the intercept parameters.
- Moderation is specified by mentioning the `"outcome ON covariate;"` syntax under each of the class-specific statements.
- Note that the binary covariate is centered so that reported distal means (intercepts) are estimated at the weighted average of `COVAR1`.

```
m_step3  <- mplusObject(
  TITLE = "Step3  (MANUAL 3-STEP ML APPROACH)",

  VARIABLE =
 "nominal = N;
  usevar = n;
  missing are all (999);

  usevar = COVAR1 DISTAL1 DISTAL2;
  classes = c(3); ",

  DEFINE =
 "Center COVAR1 (Grandmean);",

  ANALYSIS =
 "estimator = mlr;
  type = mixture;
  starts = 0;",

  MODEL =
  glue(
 "!!! OUTCOMES = DISTAL1 DISTAL2 !!!
  !!! MODERATOR = COVAR1         !!!

  %OVERALL%
  DISTAL1 on COVAR1;
  DISTAL1;

  DISTAL2 on COVAR1;
  DISTAL2;

  %C#1%
  [n#1@{logit_cprobs[1,1]}];
  [n#2@{logit_cprobs[1,2]}];

  [DISTAL1](m01);
  DISTAL1;                    !!! estimate conditional intercept !!!
  DISTAL1 on COVAR1 (s01);    !!! estimate conditional regression !!!

  [DISTAL2] (m1);
  DISTAL2;
  DISTAL2 on COVAR1 (s1);

  %C#2%
  [n#1@{logit_cprobs[2,1]}];
  [n#2@{logit_cprobs[2,2]}];

  [DISTAL1](m02);
  DISTAL1;
  DISTAL1 on COVAR1 (s02);

  [DISTAL2] (m2);
  DISTAL2;
  DISTAL2 on COVAR1 (s2);

  %C#3%
  [n#1@{logit_cprobs[3,1]}];
  [n#2@{logit_cprobs[3,2]}];

  [DISTAL1](m03);
  DISTAL1;
  DISTAL1 on COVAR1 (s03);

  [DISTAL2] (m3);
  DISTAL2;
  DISTAL2 on COVAR1 (s3);"),

  MODELCONSTRAINT =
 "New (diff12 diff13
  diff23 slope12 slope13
  slope23 ndiff12 ndiff13
  ndiff23 nslope12 nslope13
  nslope23);
```

```
  diff12 = m1-m2;   ndiff12 = m01-m02;
  diff13 = m1-m3;   ndiff13 = m01-m03;
  diff23 = m2-m3;   ndiff23 = m02-m03;
  slope12 = s1-s2;  nslope12 = s01-s02;
  slope13 = s1-s3;  nslope13 = s01-s03;
  slope23 = s2-s3;  nslope23 = s02-s03;",

  MODELTEST =
  ## NOTE: Only a single Wald test can be conducted per model run. Therefore,
  ## this example requires running separate models for each omnibus test (e.g.,
  ## 4 models; 2 outcomes and 2 slope coefficients). This can be done by
  ## commenting out all but one test and then estimating multiple versions of the model.

 "m1=m2;       !!! Distal outcome omnibus Wald test for `DISTAL2` !!!
  m2=m3;

  !s1=s2;      !!! Slope difference omnibus Wald test `DISTAL2 on COVAR1` !!!
  !s2=s3;

  !m01=m02;    !!! Distal outcome omnibus Wald test for `DISTAL1` !!!
  !m02=m03;

  !s01=s02;   !!! Slope difference omnibus Wald test for `DISTAL2 on COVAR1` !!!
  !s02=s03;",

  usevariables = colnames(savedata),
  rdata = savedata)

m_step3_fit <- mplusModeler(m_step3,
                 dataout=here("3step_mplus", "Step3_3step.dat"),
                 modelout=here("3step_mplus", "Step3_3step.inp"),
                 check=TRUE, run = TRUE, hashfilename = FALSE)
```

End of 3-Step Procedure

References:

Asparouhov, T., & Muthén, B. O. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. Structural Equation Mod- eling, 21, 329–341. http://dx.doi.org/10.1080/ 10705511.2014.915181

Hallquist, Michael N., and Joshua F. Wiley. 2018. "MplusAutomation: An R Package for FacilitatingLarge-Scale Latent Variable Analyses in Mplus." Structural Equation Modeling, 1–18. https://doi.org/10.1080/10705511.2017.1402334.

Müller, Kirill. 2017.Here: A Simpler Way to Find Your Files. https://CRAN.R-project.org/package=here.

Muthen L.K., & Muthen B.O. (1998-2017) Mplus User's Guide. Eight Edition. Los Angelos, CA: Muthen & Muthen.

R Core Team. 2019.R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.