[PSY202B] Statistical Modeling in Psychological Sciences

Ihnwhi Heo, M.Sc.

Spring 2025

Contents

Chapter 1

Introduction

Hi everyone! I'm Ihnwhi.

It is my great pleasure to be your guest lecturer for PSY202B. The theme of my lecture is statistical modeling in psychological sciences.

An essential aspect of psychological research is statistical modeling based on substantive theories. It is thus important to use statistical software for accurate modeling to reach the research conclusion. I will briefly introduce Mplus and walk you around such analytic techniques as regression analysis, path analysis, confirmatory factor analysis, structural equation modeling (Part 1), and multilevel modeling (Part 2). This GitBook is your guide such that you can easily access code for Mplus.

Are you ready? Let's get it on!

Chapter 2

Introduction to Mplus

2.1 What is it? Why called Mplus?

Mplus is a statistical modeling program that provides researchers with a flexible tool to analyze data

- Many models: regression, path analysis, factor analysis, SEM, MLM, longitudinal models, mixture model, mediation/moderation
- Many data: cross-sectional, longitudinal, single-/multilevel, observed/latent, incomplete
- Many variables: continuous, dichotomous, categorical, count
- Many estimator: maximum likelihood, weighted least squares, Bayesian

2.2 Syntax-based programming

- Commands and subcommands (https://www.statmodel.com/language.html)
- Examples of commands? (https://www.youtube.com/watch?v=XeRRtdmu23k)
  - We will be 'mostly' using TITLE, DATA, VARIABLE, ANALYSIS, MODEL, OUTPUT commands
  - But we will also be often using DEFINE, SAVEDATA, PLOT, MONTECARLO commands

2.3 Some tips when programming

1. Comments can be added with exclamation marks (!)

2. Commands should end with colon (:), and subcommands should end with semicolon (;)
3. Syntax is not case sensitive
4. Data should consist of numeric values, with no variable names
5. Data and Mplus input file should be in the same directory (like an R working directory)

- Otherwise, be sure to specify the correct directory

2.4 Some tips about model command particularly

1. Start with a path diagram
2. Think of it as specifying model parameters
3. Care to the degrees of freedom (DF)

2.5 Example: Multiple linear regression using maximum likelihood estimation

2.5.1 Model syntax

```
! Title command
TITLE: Predicting album sales using ML multiple regression

! Data command
DATA:
    ! When data and input file are in the same working directory
    FILE IS Album Sales.csv; ! Subcommands should end with ;

    ! When data and input file are in the different working directory
    ! FILE IS c:\desktop\different folder\Album Sales.csv;

! Variable command
VARIABLE:
    ! Column names (i.e., ALL variable names)
    NAMES ARE adverts sales airplay attract;

    ! Variables that will be used in our analysis
    USEVARIABLES ARE adverts sales airplay;

! Analysis command
```

```
ANALYSIS:
    ESTIMATOR IS ML; ! This is the default

! Model command
MODEL:
    ! Let's predict sales using adverts and airplay
    ! We regress sales on adverts and airplay
    sales ON adverts airplay;

    ! If you do not specify variances of and covariances between predictors
    ! degrees of freedom (DF) are not correct
    ! Variances of exogenous variable
    adverts airplay;
    ! Covariances between exogenous variable
    adverts WITH airplay;

! Output command
OUTPUT:
    TECH1 SAMPSTAT STDYX;
    ! TECH1 to understand which parameters are being estimated
    ! SAMPSTAT to check sample descriptive statistics
    ! STDYX to standardize Y (i.e., DV) and X (i.e., IV)
```

2.5.2 Part of the output file

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

 SALES    ON
    ADVERTS          0.087      0.007     12.082      0.000
    AIRPLAY          3.589      0.285     12.608      0.000

 ADVERTS  WITH
    AIRPLAY        604.061    421.412      1.433      0.152

 Means
    ADVERTS        614.412     34.255     17.936      0.000
    AIRPLAY         27.500      0.865     31.777      0.000
```

2.6 Additional materials

1. Official website at https://www.statmodel.com/

2. User's guide and examples at https://www.statmodel.com/ugexcerpts.shtml → Highly recommended!
3. Mplus YouTube channel at https://www.youtube.com/c/MplusVideos
4. QuantFish YouTube channel at https://www.youtube.com/c/QuantFish
5. Tutorials by Prof. Rens van de Schoot and his students at https://www.rensvandeschoot.com/tutorials/

Chapter 3

Path Analysis

3.1 Research scenario

A team of researchers (Lucas, Alexis, Michelle, Fei, and Marisela) is interested in understanding the impact of anxiety and distress on hostile behavior. They are thus to conduct a path analysis to examine the interrelationships between the variables. According to the substantive theory, depression (`depress`) is predicted by anxiety (`anxiety`), and hostile behavior (`hostile`) is predicted by depression and distress. So here, depression is a mediating variable between anxiety and hostility. Two exogenous variables are anxiety and distress. Use the `mmpi.csv` dataset.

3.2 Main questions

1. Draw a path diagram.
2. Write Mplus syntax.

```
! Annotate what you are doing in this line
title: Path analysis

data:
! Annotate what you are doing in this line
file is mmpi.csv;

variable:
! Annotate what you are doing in this line
names are subid source age sex race slpneed slpget anxiety depress
hostile posafect senseek totaccpt totsavoid totici tottrust totmach
totpower totsms aggress impulse harm epixtra epineuro totrathus
```

```
totfaith totcyc totsd tothypo avoid distress;

! Annotate what you are doing in this line
usevariables are anxiety depress hostile distress;

analysis:
! Annotate what you are doing in this line
estimator = ML;

model:
! Annotate what you are doing in this line
depress on anxiety;

! Annotate what you are doing in this line
hostile on depress;

! Annotate what you are doing in this line
hostile on distress;

! Annotate what you are doing in this line
anxiety;

! Annotate what you are doing in this line
distress;

! Annotate what you are doing in this line
anxiety with distress;

output:
! Annotate what you are doing in this line
TECH1;

! Annotate what you are doing in this line
stdyx;

! Annotate what you are doing in this line
modindices (3.84);
```

3. Run the analysis and interpret the results.

3.3 Bonus questions

1. Check the model fit. Are there any possibilities of improving model fit? Explore such possibilities using modification indices.

2. Calculate the degrees of freedom by hand. Compare your result with that of Mplus.

Chapter 4

Confirmatory Factor Analysis

4.1 Research scenario

A team of developmental psychologists (Michelle, Lucas, and Alexis) is interested in constructing a psychometric theory on being manipulative to others. Our substantive theory suggests that sensation seeking (`senseek`), Machiavellianism (`totmach`), powerlessness (`totpower`), social monitoring (`totsms`), and faith (`totfaith`) measure the one underlying latent construct: being manipulative. We are thus to conduct a confirmatory factor analysis that measures manipulativeness by the five measures. Use the `mmpi.csv` dataset.

4.2 Main questions

1. Draw a path diagram.
2. Write Mplus syntax.

```
! Annotate what you are doing in this line
title: Confirmatory factor analysis

data:
! Annotate what you are doing in this line
file is mmpi.csv;

variable:
! Annotate what you are doing in this line
names are subid source age sex race slpneed slpget anxiety depress
hostile posafect senseek totaccpt totsavoid totici tottrust totmach
```

```
totpower totsms aggress impulse harm epixtra epineuro totrathus
totfaith totcyc totsd tothypo avoid distress;

! Annotate what you are doing in this line
usevariables are senseek totmach totpower totsms totfaith;

analysis:
! Annotate what you are doing in this line
estimator = ML;

model:
! Annotate what you are doing in this line
manipulativeness BY senseek totmach totpower totsms totfaith;

! In case you want to free the first factor loading
! and instead scale the variance of the factor
! manipulativeness BY senseek* totmach totpower totsms totfaith;
! manipulativeness@1;

output:
! Annotate what you are doing in this line
TECH1;

! Annotate what you are doing in this line
stdyx;
```

3. Run the analysis and interpret the results.

4.3 Bonus questions

1. By default, Mplus fixes the first factor loading to 1 for model identification. What if we want to free the first factor loading and instead scale the variance of the factor?
2. Calculate the degrees of freedom by hand. Compare your result with that of Mplus.

Chapter 5

Structural Equation Modeling

5.1 Research scenario

Based on the findings on the impact of anxiety and distress (Yepez et al., 2024; Gao et al., 2025; Reimers-Contreras & Franco, 2023) and manipulative behavior (Parnell et al., 2024), Ihnwhi is interested in modeling the interrelationships between manipulativeness, anxiety, and distress. In particular, the research aim is to predict being manipulative using anxiety (`anxiety`) and distress (`distress`) via structural equation modeling. For the sake of convenience, you can only use the three following variables to measure manipulativeness: sensation seeking (`senseek`), Machiavellianism (`totmach`), and social monitoring (`totsms`). Use the `mmpi.csv` dataset.

5.2 Main questions

1. Draw a path diagram.
2. Write Mplus syntax.

```
! Annotate what you are doing in this line
title: Structural equation modeling

data:
! Annotate what you are doing in this line
file is mmpi.csv;

variable:
! Annotate what you are doing in this line
```

```
names are subid source age sex race slpneed slpget anxiety depress
hostile posafect senseek totaccpt totsavoid totici tottrust totmach
totpower totsms aggress impulse harm epixtra epineuro totrathus
totfaith totcyc totsd tothypo avoid distress;

! Annotate what you are doing in this line
usevariables are anxiety senseek totmach totsms distress;

analysis:
! Annotate what you are doing in this line
estimator = ML;

model:
! Annotate what you are doing in this line
manipulativeness BY senseek totmach totsms;

! Annotate what you are doing in this line
manipulativeness on anxiety;

! Annotate what you are doing in this line
manipulativeness on distress;

! Annotate what you are doing in this line
anxiety;

! Annotate what you are doing in this line
distress;

! Annotate what you are doing in this line
anxiety with distress;

output:
! Annotate what you are doing in this line
TECH1;

! Annotate what you are doing in this line
stdyx;

! Annotate what you are doing in this line
modindices (3.84);
```

3. Run the analysis and interpret the results.

5.3 Bonus questions

1. Check the model fit. Are there any possibilities of improving model fit? Explore such possibilities using modification indices.
2. Can you interpret the TECH1 output given that the model formulation in Mplus is based on the LISREL "all-y" notation?

Chapter 6

Multilevel Modeling

6.1 Multilevel data

We have simulated data from 100 classes, with a different number of pupils in each class. The average class size is 20 pupils. On the pupil level, we have two variables. First is the dependent variable 'popularity', measured on a self-rating scale that ranges from 0 (very unpopular) to 10 (very popular). Second is the independent variable 'extraversion', measured on a self-rating scale ranging from 1 to 10. On the class level, we have one explanatory variable 'teacher experience', measured in years ranging from 2 to 25.

6.2 Building the multilevel regression model

We are to build three multilevel regression models in this practical. The three models to be built are as follows:

- Empty model (aka. intercept-only model)
- Model with a level-1 predictor
- Model with a level-2 predictor

For simplicity and illustrative purposes, we only consider random intercepts but not random slopes. On Wednesday, you will build the same three models in HLM software using a different dataset.

6.3 Main questions

6.3.1 Empty model (aka. intercept-only model)

1. Write Mplus syntax.

```
! Annotate what you are doing in this line
TITLE: Empty model (intercept-only model)

DATA:
    ! Annotate what you are doing in this line
    file is popular2.dat;

VARIABLE:
    ! Annotate what you are doing in this line
    names are class pupil cons extrav sex texp popular popteach zextrav
    zsex ztexp zpopular zpoptch;

    ! Annotate what you are doing in this line
    usevariables are popular;

    ! Annotate what you are doing in this line
    cluster is class;

ANALYSIS:
    ! Annotate what you are doing in this line
    type is twolevel;

    ! Annotate what you are doing in this line
    estimator is MLR;

MODEL:
    ! Annotate what you are doing in this line
    %within%

    ! Annotate what you are doing in this line
    %between%

OUTPUT:
    ! Annotate what you are doing in this line
    sampstat cinterval;
```

2. Run the analysis and interpret the results.

3. Can you match the parameter estimates to the notation in the formula below?

Level 1

$$y_{ij} = \beta_{0j} + r_{ij}$$

Level 2

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

4. Can you compute the ICC?

6.3.2 Model with a level-1 predictor

1. Write Mplus syntax.

```
! Annotate what you are doing in this line
TITLE: Model with a level-1 predictor

DATA:
    ! Annotate what you are doing in this line
    file is popular2.dat;

VARIABLE:
    ! Annotate what you are doing in this line
    names are class pupil cons extrav sex texp popular popteach zextrav
    zsex ztexp zpopular zpoptch;

    ! Annotate what you are doing in this line
    usevariables are extrav popular;

    ! Annotate what you are doing in this line
    cluster is class;

    ! Annotate what you are doing in this line
    within are extrav;

ANALYSIS:
    ! Annotate what you are doing in this line
    type is twolevel;

    ! Annotate what you are doing in this line
    estimator is MLR;

MODEL:
    ! Annotate what you are doing in this line
    %within%
    popular on extrav;
```

```
    ! Annotate what you are doing in this line
    %between%
    popular;

OUTPUT:
    ! Annotate what you are doing in this line
    sampstat cinterval;

SAVEDATA:
    ! Annotate what you are doing in this line
    file is fscores.dat;

    ! Annotate what you are doing in this line
    save is fscores;
```

2. Run the analysis and interpret the results.
3. Can you match the parameter estimates to the notation in the formula below?

Level 1

$$y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$$

Level 2

$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$

4. Can you compute the ICC?

6.3.3 Model with a level-2 predictor

1. Write Mplus syntax.

```
! Annotate what you are doing in this line
TITLE: Model with a level-2 predictor

DATA:
    ! Annotate what you are doing in this line
    file is popular2.dat;

VARIABLE:
    ! Annotate what you are doing in this line
    names are class pupil cons extrav sex texp popular popteach zextrav
    zsex ztexp zpopular zpoptch;

    ! Annotate what you are doing in this line
```

```
    usevariables are texp popular;

    ! Annotate what you are doing in this line
    cluster is class;

    ! Annotate what you are doing in this line
    between are texp;

ANALYSIS:
    ! Annotate what you are doing in this line
    type is twolevel;

    ! Annotate what you are doing in this line
    estimator is MLR;

MODEL:
    ! Annotate what you are doing in this line
    %within%
    popular;

    ! Annotate what you are doing in this line
    %between%
    popular on texp;

OUTPUT:
    ! Annotate what you are doing in this line
    sampstat cinterval;
```

2. Run the analysis and interpret the results.
3. Can you match the parameter estimates to the notation in the formula below?

Level 1

$$y_{ij} = \beta_{0j} + r_{ij}$$

Level 2

$$\beta_{0j} = \gamma_{00} + \gamma_{01} X_j + u_{0j}$$

4. Can you compute the ICC?

6.4 Bonus questions

1. Can you visualize the results of multilevel modeling? Let's practice with the model with a level-1 predictor that we fitted.

```r
# Clean the work space
rm(list=ls()); gc()
```

```
##          used (Mb) gc trigger (Mb) limit (Mb) max used (Mb)
## Ncells 501216 26.8    1097498 58.7         NA   669094 35.8
## Vcells 912551  7.0    8388608 64.0      16384  1839769 14.1
```

```r
# Load required packages
library(MplusAutomation)
```

```
## Warning: package 'MplusAutomation' was built under R version 4.2.3
```

```
## Version:  1.1.1
## We work hard to write this free software. Please help us get credit by citing:
##
## Hallquist, M. N. & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitati
##
## -- see citation("MplusAutomation").
```

```r
library(texreg)
```

```
## Version:  1.39.4
## Date:     2024-07-23
## Author:   Philip Leifeld (University of Manchester)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

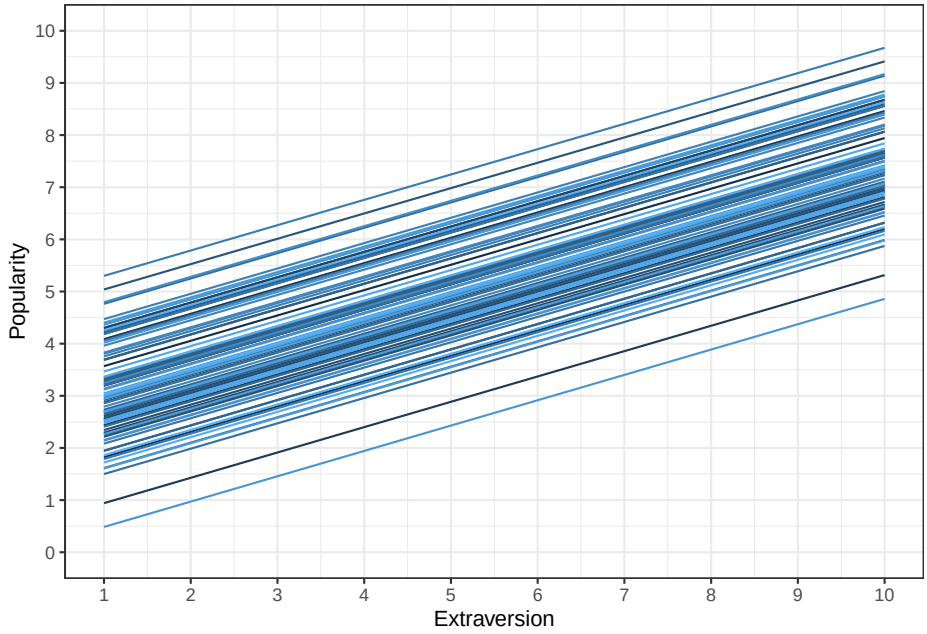
```r
# Extract fscores
out <- readModels("2. Model with a l1 pred.out", recursive = FALSE,
                  what = "savedata")
fscores <- out$savedata
fscores_classid <- aggregate(fscores[,3:4], list(fscores$CLASS), mean)

# Check the range of the predictor
range(fscores$EXTRAV) # 1 to 10
```

```
## [1]  1 10
```

```r
# Find predicted values for each level of extraversion
pred_popular_classid <- data.frame(fscores_classid[rep(seq(nrow(fscores_classid)),
                                                       each = 10),],
                                   "extraversion" = 1:10)
pred_popular <- pred_popular_classid$B_POPULAR + 0.486 * pred_popular_classid$extravers
pred_popular_classid <- data.frame(pred_popular_classid, pred_popular)
```

```
# Visualization
ggplot(data = pred_popular_classid, aes(x = extraversion, y = pred_popular, group = Group.1)) +
  geom_line(aes(color = Group.1), show.legend = FALSE) +
  labs(x = "Extraversion",
       y = "Popularity") +
  scale_y_continuous(lim = c(0, 10), breaks = seq(0, 10, by = 1)) +
  scale_x_continuous(lim = c(1, 10), breaks = seq(1, 10, by = 1)) +
  theme_bw()
```



2. How can you interpret the plot?
3. Can you imagine how the plot would look like when we consider random slopes?