An Introduction to Mplus and Path Analysis

PSYC 943: Fundamentals of Multivariate Modeling

Lecture 7: October 9, 2013

Today's Lecture

- A brief intro to Mplus

- Path analysis
  - …starting with multivariate regression…
  - …then arriving at our final destination

- Path analysis details:
  - Standardized coefficients (introduced in regression)
  - Model fit (introduced in multivariate regression)
  - Model modification (introduced in multivariate regression and path analysis)

- Additional issues in path analysis
  - Estimation types
  - Variable considerations

Today's Data Example

- Data are simulated based on the results reported in:

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: a path analysis. *Journal of Educational Psychology, 86*, 193-203.

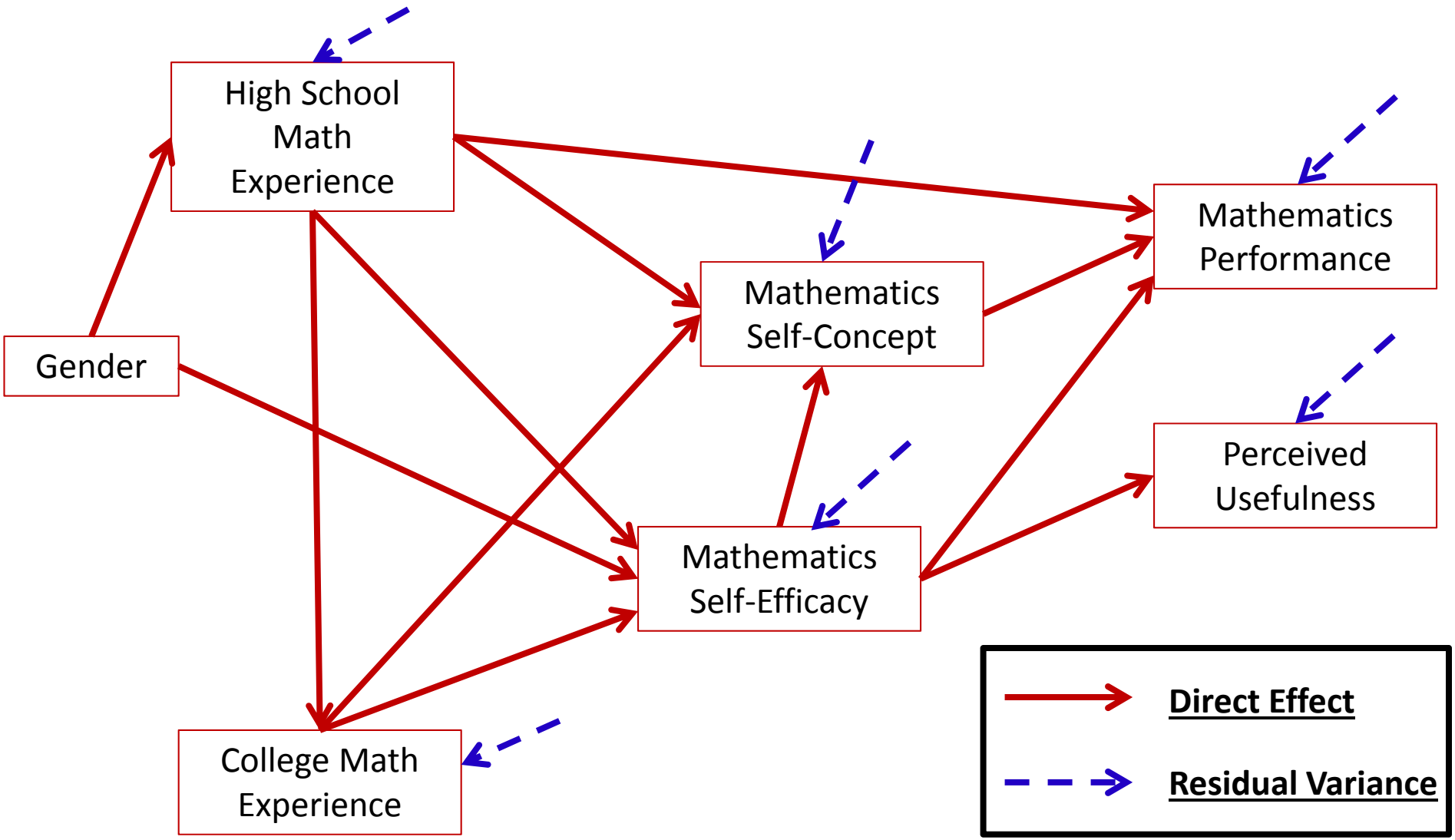
- Sample of 350 undergraduates (229 women, 121 men)
  - In simulation, 10% of variables were missing (using missing completely at random mechanism)

- Note: simulated data characteristics differ from actual data (some variables extend beyond their official range)
  - Simulated using Multivariate Normal Distribution
    - Some variables had boundaries that simulated data exceeded
  - Results will not match exactly due to missing data and boundaries

Variables of Data Example

- Gender (1 = male; 0 = female)
- Math Self-Efficacy (MSE)
  - Reported reliability of .91
  - Assesses math confidence of college students
- Perceived Usefulness of Mathematics (USE)
  - Reported reliability of .93
- Math Anxiety (MAS)
  - Reported reliability ranging from .86 to .90
- Math Self-Concept (MSC)
  - Reported reliability of .93 to .95
- Prior Experience at High School Level (HSL)
  - Self report of number of years of high school during which students took mathematics courses
- Prior Experience at College Level (CC)
  - Self report of courses taken at college level
- Math Performance (PERF)
  - Reported reliability of .788
  - 18-item multiple choice instrument (total of correct responses)

Our Destination: Overall Path Model



The Big Picture

- Path analysis is a multivariate statistical method that, when using an identity link, assumes the variables in an analysis are multivariate normally distributed
  - Mean vectors
  - Covariance matrices

- By specifying simultaneous regression equations (the core of path models), a very specific covariance matrix is implied
  - This is where things deviate from our familiar R matrix

- Like multivariate models, the key to path analysis is finding an approximation to the unstructured (saturated) covariance matrix
  - With fewer parameters, if possible

- The art to path analysis is in specifying models that blend theory and statistical evidence to produce valid, generalizable results

INTRODUCTION TO MPLUS

The Mplus Statistical Package

- Mplus provides a general latent variable modeling framework that allows for combinations of:
  - Continuous or categorical latent variables
  - Continuous, categorical, count, nominal or censored data

- Mplus is commercial software that is available on Windows machines in the Burnett computer labs

- Mplus is also available for purchase:
  - Available at http://www.statmodel.com
  - From $195 to $350 (student)

Mplus Data File Input Format

- Mplus input files must be ASCII text based (so not binary)
  - Text-based file formats: *.txt, *.dat, *.csv
  - Not-text-based file formats: *.xlsx, *.sas7bdat, *.sav

- The easiest way to get data files into Mplus is to use "free-formatting" (some type of delimiter between columns)
  - I prefer comma-delimited files and will only use those in this course
  - **Cannot start with variable names in first row of data**

- Typically, I store data in Excel and save as a comma-delimited file
  - Save As…*.csv…
    - (then click OK to the first question)…(then click YES to the second)
    - Ignore the warning (click NO) to re-save when closing the Excel Workbook

Mplus Syntax Conventions

- Most syntax must have a semi-colon end each line (;)
  - Exceptions: TITLE section, comments, and continuing lines
- Comments are denoted with an exclamation point (!)
- Syntax is organized by sections; headings of sections end with colons (:)
  - TITLE:, DATA:, VARIABLE:, DEFINE:, and MODEL: are what we use this week
- Syntax cannot exceed 90 characters per row (ƾ)
- Mplus input files are typically saved with the extension *.inp
- Mplus output files are typically saved with the extension *.out
  - Both are ASCII text (i.e., you can open with text editors)
- The default location for the data file and output file are the folder containing the input file

Mplus TITLE Section

- The TITLE section contains the label of the analysis
- You can type whatever you want here…it will appear verbatim at the top of the output file
- You do not have to terminate this section with a semi-colon
- This section is optional

Mplus DATA Section

- The DATA section is where data files are defined

- Define the name (and path if different from input file folder) by using the command:

FILE = mydata.csv

- POTENTIAL MISTAKES:
  - Data must be numeric – if not errors happen
  - First row of data should not contain variable names

- This section is NOT OPTIONAL

Mplus VARIABLE Section

- The Mplus VARIABLE section defines the names of the variables in the data file, variable types, and variables in your analysis

  - NAMES = provides variable names
    - Names cannot be more than 8 characters
    - Lists of variables can be created (i.e., X1-X10 makes 10 variables)
    - By default all variables listed in the NAMES section are assumed to be part of the analysis
  - USEVARIABLE = provides names of variables used in the analysis (optional)
  - IDVARIABLE = provides the ID variable name (optional)

- This section is NOT OPTIONAL

Mplus DEFINE Section

- The DEFINE section is where new variables are created

- To test our equal slopes hypothesis we created interaction variables by the following syntax:

```
TITLE: !title section puts text below at the top of the output file
    ANCOVA MODEL WITH GENDER, MOTIVATION, AND ACHIEVEMENT
    TESTING INTERACTIONS - DIFFERENT SLOPES WITHIN GENDER GROUP
DATA: !data section defines data file
    FILE = exampledata.csv;

VARIABLE: !variable section defines varibles in data file
    NAMES = ID achieve selfest motivate gender se1-se5
            motiv1-motiv5 ach1-ach20;
    IDVARIABLE = ID;
    USEVARIABLE = achieve selfest gender motivate femaleSE femaleM;

DEFINE:
    femaleSE = gender*selfest;
    femaleM = gender*motivate;

MODEL:
    achieve ON selfest gender motivate femaleSE femaleM;
```

Mplus MODEL Section

- The MODEL section is where you define the model

  These models use the ON statement (ON = REGRESSION)

  achieve ON selfest motivate

- And an empty model to figure out the covariance matrix of the MOTIVATION items
  - This used the WITH statement (WITH = COVARIANCE)

  motiv1-motive5 WITH motive1-motive5

LINEAR REGRESSION: A BASIC PATH MODEL

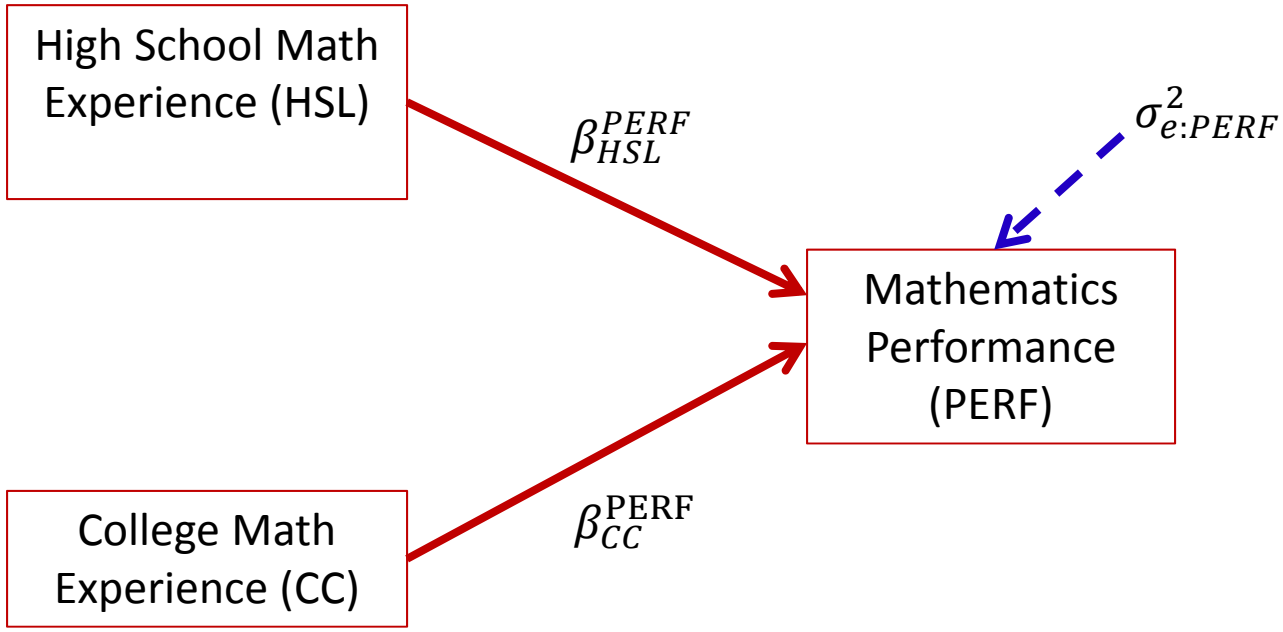
Linear Regression Framed As A Path Model

- We will begin our discussion by starting with linear regression: predicting mathematics performance (PERF) with high school (HSL) and college experience (CC)

$$PERF_i = \beta_0^{PERF} + \beta_{HSL}^{PERF} HSL_i + \beta_{CC}^{PERF} CC_i + e_i^{PERF}$$

- As typical, we assume $e_i^{PERF} \sim N(0, \sigma_{e:PERF}^2)$
- A guide to my notation:
  - β_X^Y - the regression slope where variable Y is being predicted by variable X
  - β_0^Y - the intercept for the regression line predicting variable Y
  - e_i^Y - the residual for variable Y for observation i
  - $\sigma_{e:Y}^2$ - the **residual** variance (note the e: in the subscript) for the prediction of variable Y
  - σ_X^2 - the variance of variable X (not a residual – unexplained)

Linear Regression Path Diagram



Types of Variables in the Analysis

- An important distinction in path analysis is between endogenous and exogenous variables

- Endogenous variable(s): variables whose variability ***is explained*** by one or more variables in a model
  - In linear regression, the **dependent variable** is the only endogenous variable in an analysis
    - Mathematics Performance (PERF) in our example

- Exogenous variable(s): variables whose variability ***is not explained*** by any variables in a model
  - In linear regression, the **independent variable**(s) are the exogenous variables in the analysis
    - High school (HSL) and college (CC) experience

Linear Regression in Mplus

- The basic code for linear regression in Mplus uses the ON statement:

```
VARIABLE:
    NAMES = id gender hsl cc use msc mas mse perf;
    USEVARIABLE = hsl perf cc;
    IDVARIABLE = id;
    MISSING = .;

ANALYSIS:
    ESTIMATOR = MLR;

MODEL:
    perf ON hsl cc;

OUTPUT:
    STANDARDIZED RESIDUAL;
```

- Mplus uses ML by default to estimate the parameters of the model
  - Listwise deletion happens for any independent variables (right of ON) with missing data

```
*** WARNING
  Data set contains cases with missing on x-variables.
  These cases were not included in the analysis.
  Number of cases with missing on x-variables:  68
```

- Sample should be 350 subjects
  - Mplus uses 237

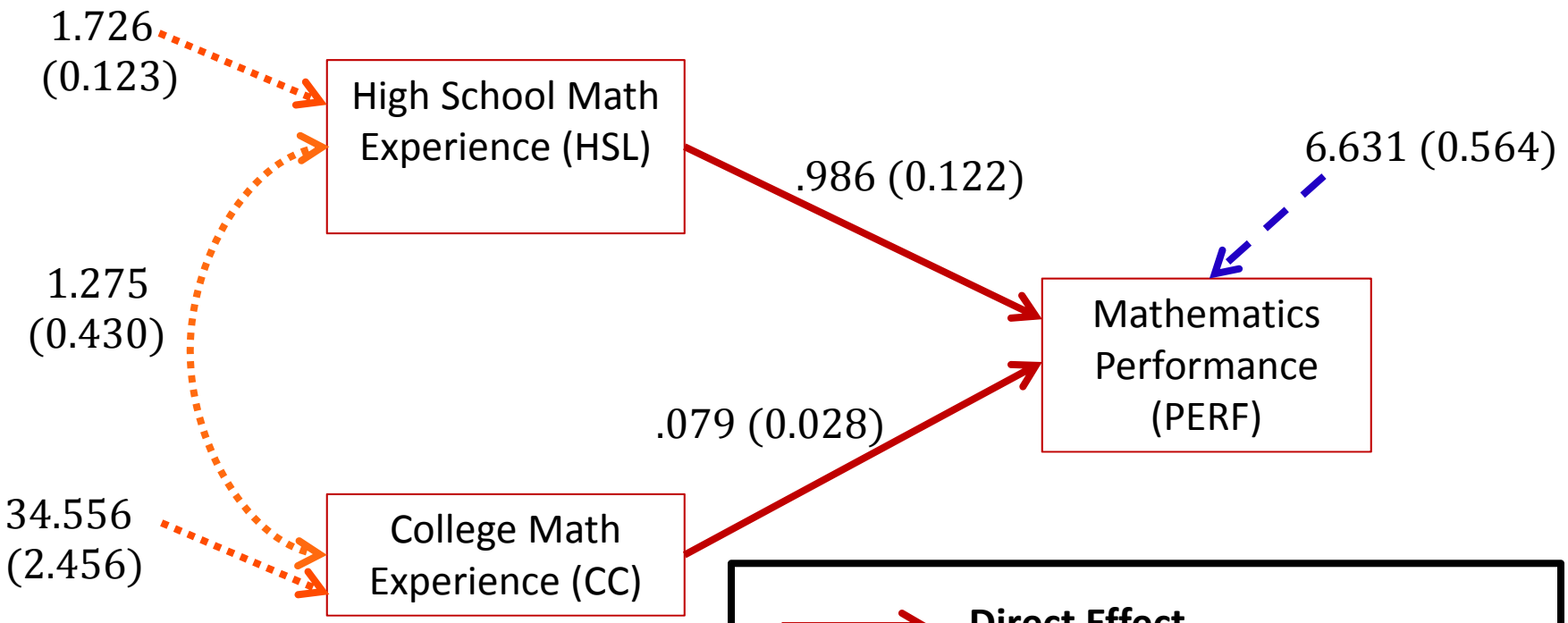
```
SUMMARY OF ANALYSIS

Number of groups                                         1
Number of observations                                 237

Number of dependent variables                            1
Number of independent variables                          2
Number of continuous latent variables                    0
```

Linear Regression Results in Mplus

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF ON				
HSL	0.986	0.122	8.103	0.000
CC	0.079	0.028	2.861	0.004
HSL WITH				
CC	1.275	0.430	2.967	0.003
Means				
HSL	4.925	0.073	67.051	0.000
CC	10.331	0.331	31.189	0.000
Intercepts				
PERF	8.253	0.598	13.807	0.000
Variances				
HSL	1.726	0.123	14.039	0.000
CC	34.556	2.456	14.069	0.000
Residual Variances				
PERF	6.631	0.564	11.759	0.000

Linear Regression Path Diagram with Results



Not Shown On Path Diagram:

- $\beta_0^{PERF} = 8.253\ (0.598)$
- $\mu_{HSL} = 4.925\ (0.073)$
- $\mu_{CC} = 10.331\ (0.331)$

Interpreting Linear Regression Results

- The linear regression results are interpreted as follows:
  - $\beta_0^{PERF} = 8.253$: the intercept for PERF – the value of PERF when all predictors are zero (HSL = 0 and CC = 0)
  - $\beta_{HSL}^{PERF} = 0.986$: the slope for HSL. Indicates that for every one-unit increase in HSL (holding CC constant), PERF increases by .986
  - $\beta_{CC}^{PERF} = 0.079$: the slope for CC. Indicates that for every one-unit increase in CC (holding HSL constant), PERF increases by .079
  - $\sigma_{e:PERF}^2 = 6.631$: the residual (or unexplained) variance in PERF
  - Note: the rest of the parameters are the descriptive statistics for the independent (exogenous) variables and are not explained by the regression model

Explained Variance

- To demonstrate the concept of explained variance, consider the dependent variable, math performance
  - "Empty Model" – estimate of its variance: $\sigma^2_{PERF} = 8.722$

ESTIMATED SAMPLE STATISTICS

Means

	PERF	HSL	CC
1	13.923	4.925	10.331

Covariances

	PERF	HSL	CC
PERF	8.722		
HSL	1.802	1.726	
CC	3.980	1.275	34.556

Note: the sample statistics are from the unstructured (saturated) model estimated with ML – if you have missing data or are using unbiased estimates, these will not match other programs

- The independent (exogenous) variables in the analysis seek to explain the variability in math performance
  - Adding significant IVs will reduce the variance, therefore "explaining" a portion of the DV

Regression Model Explained Variance

- After adding both independent variables HSL and CC, the residual variance of performance was $\sigma^2_{e:PERF} = 6.631$

- Therefore, the inclusion of these variables reduced the variance of PERF from 8.722 to 6.631, for an

$$R^2 = \frac{8.722 - 6.631}{8.722} = .24$$

- Mplus reports this value under the standardized coefficients output (explained next):

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF	0.240	0.043	5.566	0.000

Standardized Coefficients

- The scale of the (unstandardized) slope coefficients is given in terms of UNITS of Y (SD Y) per UNITS of X (SD X)
  - Y goes up β_X^Y UNITS of Y for every UNIT of X
    - HSL has SD of 1.31; CC has SD of 5.88
  - If the UNITS of X differ for the various IVs in a model, it can be hard to compare relative strengths of coefficients
    - $\beta_{HSL}^{PERF} = .986$ (but HSL has SD of 1.31)
    - $\beta_{CC}^{PERF} = .079$ (but CC has SD of 5.88)

- Standardized coefficients are the coefficients that would be obtained if Y and X were standardized:
  - Standardized = variance of 1 (i.e. z-scores used for analysis)

- Standardized coefficients are useful for comparing the relative effects of each IV in the model

Standardization in Mplus

- Under the output section, the word STANDARDIZED will produce standardized coefficients in Mplus output

- Three types of standardizations are given:
  - **STDYX**: These are the standardized regression coefficients; use these for continuous IVs (used for our current analysis)
  - **STDY**: These only standardize based on variance of Y (the DV). Use when binary variables are IVs (like gender dummy coding) as unit of X has no meaning
  - **STD:** Discussed when we get to models with latent variables

Standardized Coefficients Output

- Standardized Coef:

$$b_{effect} = \beta_{effect} \frac{SD(X_{effect})}{SD(Y)}$$

- For HSL:

$$b_{HSL}^{PERF} = .986 \frac{1.726}{8.722} = .439$$

  - PERF increases .439 SD when HSL increases 1 SD (holding CC constant)

- For CC:

$$b_{CC}^{PERF} = .079 \frac{34.556}{8.722} = .157$$

  - PERF increases .157 SD when CC increases 1 SD (holding HSL constant)

STDYX Standardization

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF	ON				
HSL		0.439	0.047	9.415	0.000
CC		0.157	0.055	2.875	0.004

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF	ON				
HSL		0.986	0.122	8.103	0.000
CC		0.079	0.028	2.861	0.004
Variances					
HSL		1.726	0.123	14.039	0.000
CC		34.556	2.456	14.069	0.000

RESIDUAL OUTPUT

Model Estimated Covariances/Correlations/Residual Correlations

	PERF	HSL	CC
PERF	8.722		
HSL	1.802	1.726	
CC	3.980	1.275	34.556

MULTIVARIATE REGRESSION

Multivariate Regression

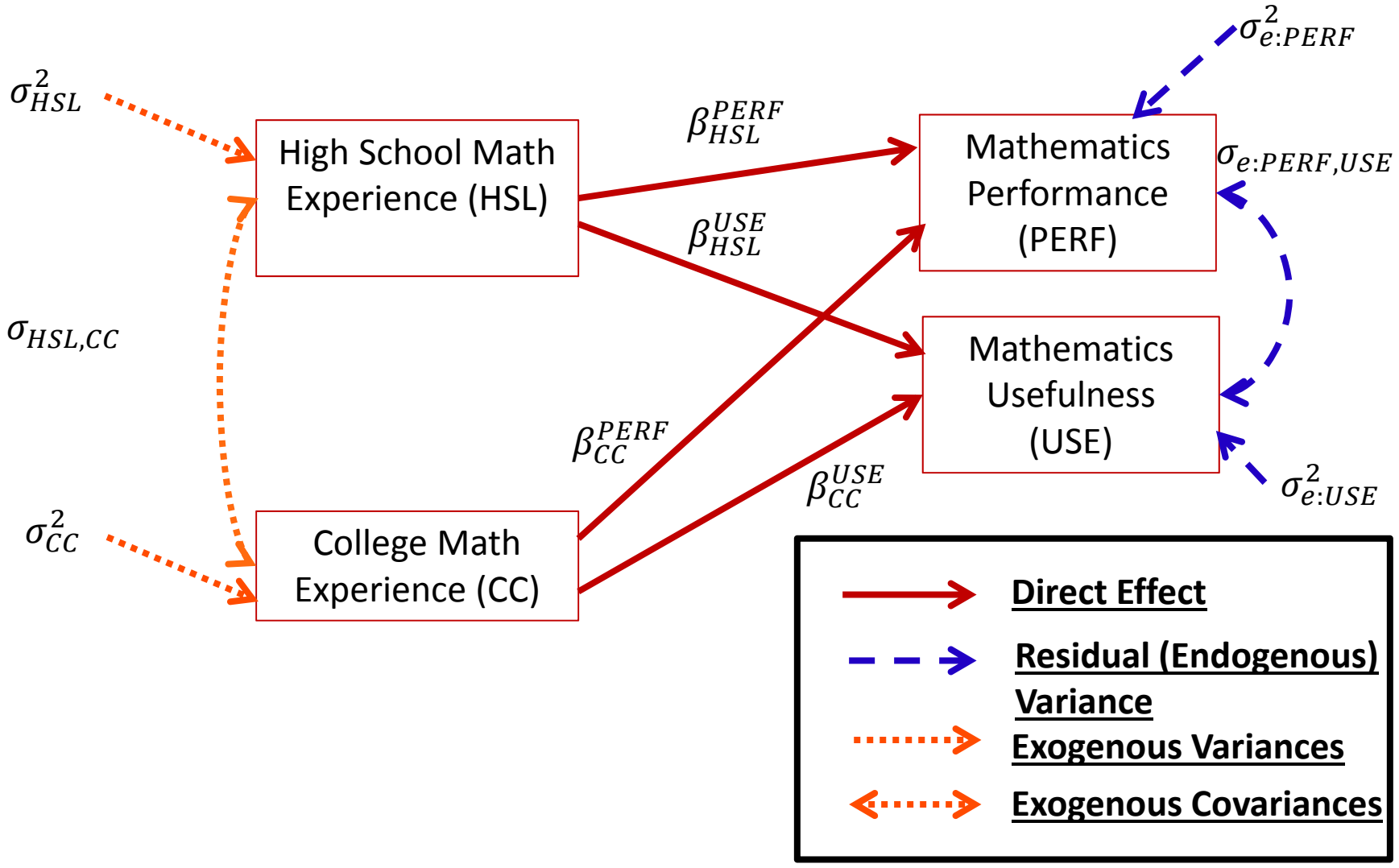
- Before we dive into path analysis, we will begin with a multivariate regression model:
  - Predicting mathematics performance (PERF) with high school (HSL) and college (CC) experience
  - Predicting perceived usefulness (USE) with high school (HSL) and College (CC) experience

$$PERF_i = \beta_0^{PERF} + \beta_{HSL}^{PERF} HSL_i + \beta_{CC}^{PERF} CC_i + e_i^{PERF}$$
$$USE_i = \beta_0^{USE} + \beta_{HSL}^{USE} HSL_i + \beta_{CC}^{USE} CC_i + e_i^{USE}$$

- We denote the residual for PERF as e_i^{PERF} and the residual for USE as e_i^{USE}
  - We also assume the residuals are Multivariate Normal:

$$\begin{bmatrix} e_i^{PERF} \\ e_i^{USE} \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e:PERF}^2 & \sigma_{e:PERF,USE} \\ \sigma_{e:PERF,USE} & \sigma_{e:USE}^2 \end{bmatrix} \right)$$

Multivariate Linear Regression Path Diagram



Types of Variables in the Analysis

- An important distinction in path analysis is between endogenous and exogenous variables

- Endogenous variable(s): variables whose variability ***is explained*** by one or more variables in a model
  - In linear regression, the **dependent variable** is the only endogenous variable in an analysis
    - Mathematics Performance (PERF) and Mathematics Usefulness (USE)

- Exogenous variable(s): variables whose variability ***is not explained*** by any variables in a model
  - In linear regression, the **independent variable**(s) are the exogenous variables in the analysis
    - High school (HSL) and college (CC) experience

Multivariate Regression in Mplus

- The basic code for linear regression in Mplus uses the ON statement
- The WITH statement estimates a covariance between the two variables

```
TITLE:
    MULTIVARIATE Multiple Regression Analysis
    Predicting Performance and Perceived Usefulness
    NOTE: NO LISTWISE DELETION OF INCOMPLETE CASES
    THIS IS DUE TO ADDING THE WITH STATEMENT (INSERTS INTO LIKELIHOOD FUNCTION)

DATA:
    FILE = mathdata.csv;

VARIABLE:
    NAMES = id gender hsl cc use msc mas mse perf;
    USEVARIABLE = hsl perf cc use;
    IDVARIABLE = id;
    MISSING = .;

MODEL:
    perf ON hsl cc;
    use ON hsl cc;
    hsl; cc; hsl WITH cc; !adds exogenous variables to likelihood

OUTPUT:
    STANDARDIZED RESIDUAL SAMPSTAT;
```

Labeling Variables

- The endogenous (dependent) variables are:
  - Performance (PERF) and Usefulness (USE)

- The exogenous (independent) variables are:
  - High school (HSL) and college (CC) experience

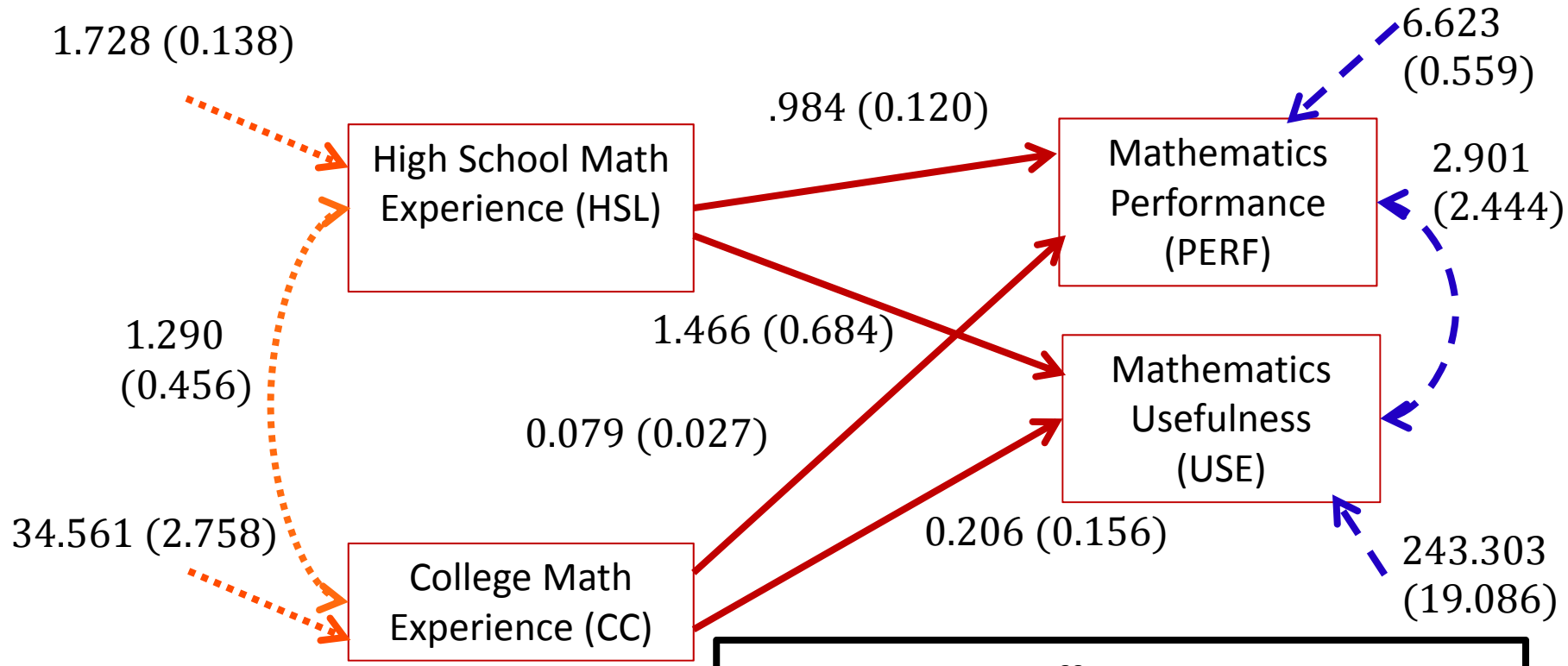
Multivariate Regression Model Parameters

- If we considered all four variables to be part of a multivariate normal distribution, our unstructured (saturated) model would have a total of 14 parameters:
  - 4 means
  - 4 variances
  - 6 covariances (4-choose-2 or 4*(4-1)/2))

- The model itself has 14 parameters:
  - 4 intercepts
  - 4 slopes
  - 2 residual variances
  - 1 residual covariance
  - 2 exogenous variances
  - 1 exogenous covariance

- Therefore, this model will fit perfectly – no model fit statistics will be available
  - Even without model fit, interpretation of parameters can proceed

Multivariate Linear Regression Path Diagram (Unstandardized Coefficients)



Not Shown On Path Diagram:

- $\beta_0^{PERF} = 8.264\ (0.629)$
- $\beta_0^{USE} = 43.129\ (0.359)$
- $\mu_{HSL} = 4.922\ (0.074)$
- $\mu_{CC} = 10.330\ (0.331)$

Interpreting Multivariate Regression Results for PERF

- $\beta_0^{PERF} = 8.264$: the intercept for PERF – the value of PERF when all predictors are zero (HSL = 0 and CC = 0)

- $\beta_{HSL}^{PERF} = 0.986$: the slope for HSL predicting PERF. Indicates that for every one-unit increase in HSL (holding CC constant), PERF increases by .986
  - The standardized coefficient was .438

- $\beta_{CC}^{PERF} = 0.079$: the slope for CC predicting PERF. Indicates that for every one-unit increase in CC (holding HSL constant), PERF increases by .079
  - The standardized coefficient was .157

Interpreting Multivariate Regression Results for USE

- $\beta_0^{USE} = 43.129$: the intercept for USE – the value of USE when all predictors are zero (HSL = 0 and CC = 0)

- $\beta_{HSL}^{USE} = 1.466$: the slope for HSL predicting USE. Indicates that for every one-unit increase in HSL (holding CC constant), USE increases by 1.466
  - The standardized coefficient was .122

- $\beta_{CC}^{USE} = 0.206$: the slope for CC predicting USE. Indicates that for every one-unit increase in CC (holding HSL constant), USE increases by .206. This was found to be not significant, meaning college experience did not predict perceived usefulness
  - The standardized coefficient was .077

Interpretation of Residual Variances and Covariances

- $\sigma^2_{e:PERF} = 6.623$: the residual variance for PERF
  - The R^2 for PERF was .240 (the same as before)

- $\sigma^2_{e:USE} = 243.303$: the residual variance for USE
  - The R^2 for USE was .024 (a very small effect)

- $\sigma_{e:PERF,USE} = 2.901$: the residual covariance between USE and PERF
  - This value was not significant, meaning we can potentially set its value to zero and re-estimate the model

- Each of these variance describes the amount of variance not accounted for in each dependent (endogenous) variable

Overall Model R^2 for All Endogenous Variables

- Although the residual variance and R^2 values for PERF and USE describe how each variable is explained individually, we can use multivariate statistics to describe the joint explanation of both
  - R^2 comparing the generalized variances (determinant of covariance matrix)

- The overall generalized variance of the endogenous variables without the model was $|\boldsymbol{\Sigma}| = \left|\begin{bmatrix} 8.709 & 6.362 \\ 6.362 & 249.254 \end{bmatrix}\right| = 2{,}130.28$

- The generalized **residual** variance of the endogenous variables was $|\widehat{\boldsymbol{\Sigma}}| = \left|\begin{bmatrix} 6.623 & 2.901 \\ 2.901 & 243.303 \end{bmatrix}\right| = 1{,}602.98$

- Therefore, the generalized R^2 was $\frac{2{,}130.28 - 1{,}602.98}{2{,}130.28} = .248$
  - Most of that came from the PERF variable

Comparison of Model Output from Linear and Multivariate Regression Models

Linear Regression

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF	ON				
HSL		0.986	0.120	8.191	0.000
CC		0.079	0.027	2.930	0.003
HSL	WITH				
CC		1.275	0.456	2.796	0.005
Means					
HSL		4.925	0.073	67.022	0.000
CC		10.331	0.331	31.170	0.000
Intercepts					
PERF		8.253	0.631	13.084	0.000
Variances					
HSL		1.726	0.137	12.573	0.000
CC		34.556	2.757	12.534	0.000
Residual Variances					
PERF		6.631	0.560	11.841	0.000

- Results for linear regression parameters will be virtually unchanged
- Here, they differ due to one extra observation included in model

Multivariate Regression

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF	ON				
HSL		0.984	0.120	8.183	0.000
CC		0.079	0.027	2.934	0.003
USE	ON				
HSL		1.466	0.684	2.143	0.032
CC		0.206	0.156	1.317	0.188
HSL	WITH				
CC		1.290	0.456	2.827	0.005
USE	WITH				
PERF		2.901	2.444	1.187	0.235
Means					
HSL		4.922	0.074	66.952	0.000
CC		10.330	0.331	31.174	0.000
Intercepts					
PERF		8.264	0.629	13.129	0.000
USE		43.129	3.590	12.014	0.000
Variances					
HSL		1.728	0.138	12.562	0.000
CC		34.561	2.758	12.533	0.000
Residual Variances					
PERF		6.623	0.559	11.846	0.000
USE		243.303	19.086	12.748	0.000

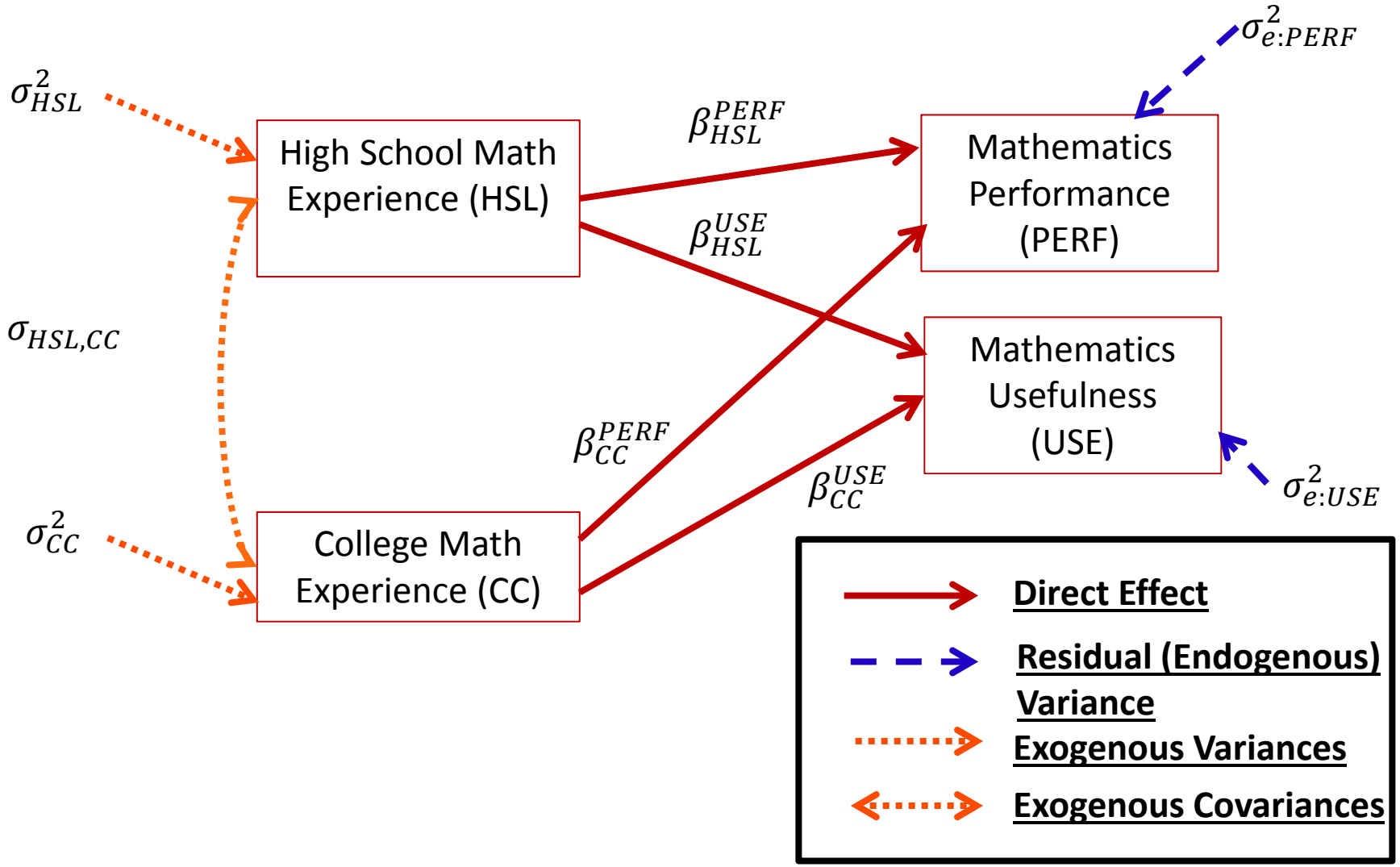
Model Modification

- The residual covariance parameter (between PERF and USE) was not significant
- This means that after accounting for the relationship between HSL and CC with PERF along with HSL and CC with USE, the correlation between these two is zero
  - Meaning we can likely remove the parameter from the model

```
MODEL:
    perf ON hsl cc;
    use ON hsl cc;
    hsl WITH cc;

    perf WITH use @0;
```

- Removal of the parameter from the model would reduce the number of estimated parameters from 14 to 13
  - And would provide a mechanism to inspect goodness of fit of the reduced model

Reduced Model Path Diagram



COMPARING SAS AND MPLUS

In Class Demonstration using Example SAS and Mplus Files

- See video…

WRAPPING UP

Wrapping Up

- Today's lecture was meant to provide a visual schematic of path diagrams to help with the multivariate modeling section of the course

- Over the next few weeks, we will investigate more multivariate models – but continue to use path diagrams to help illustrate each of the effects we seek to find

- Our lectures will culminate with the full path diagram used from this example – when multivariate models feature variables that are both predictors and dependent variables
  - SAS PROC MIXED cannot estimate these models