Multivariate Linear Models with Predictors

EPSY 905: Fundamentals of Multivariate Modeling

Online Lecture #13

In This Lecture

- Multivariate linear models with predictors (using path analysis software/packages)

- Details and terminology from path analysis:
  - Variable naming conventions
  - Software estimation defaults (variables in/out of likelihood)
  - Model comparisons via likelihood ratio tests
  - Measures of absolute and approximate model fit
  - Model modification methods
  - Standardized regression coefficients

- Additional issues in path analysis
  - 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.

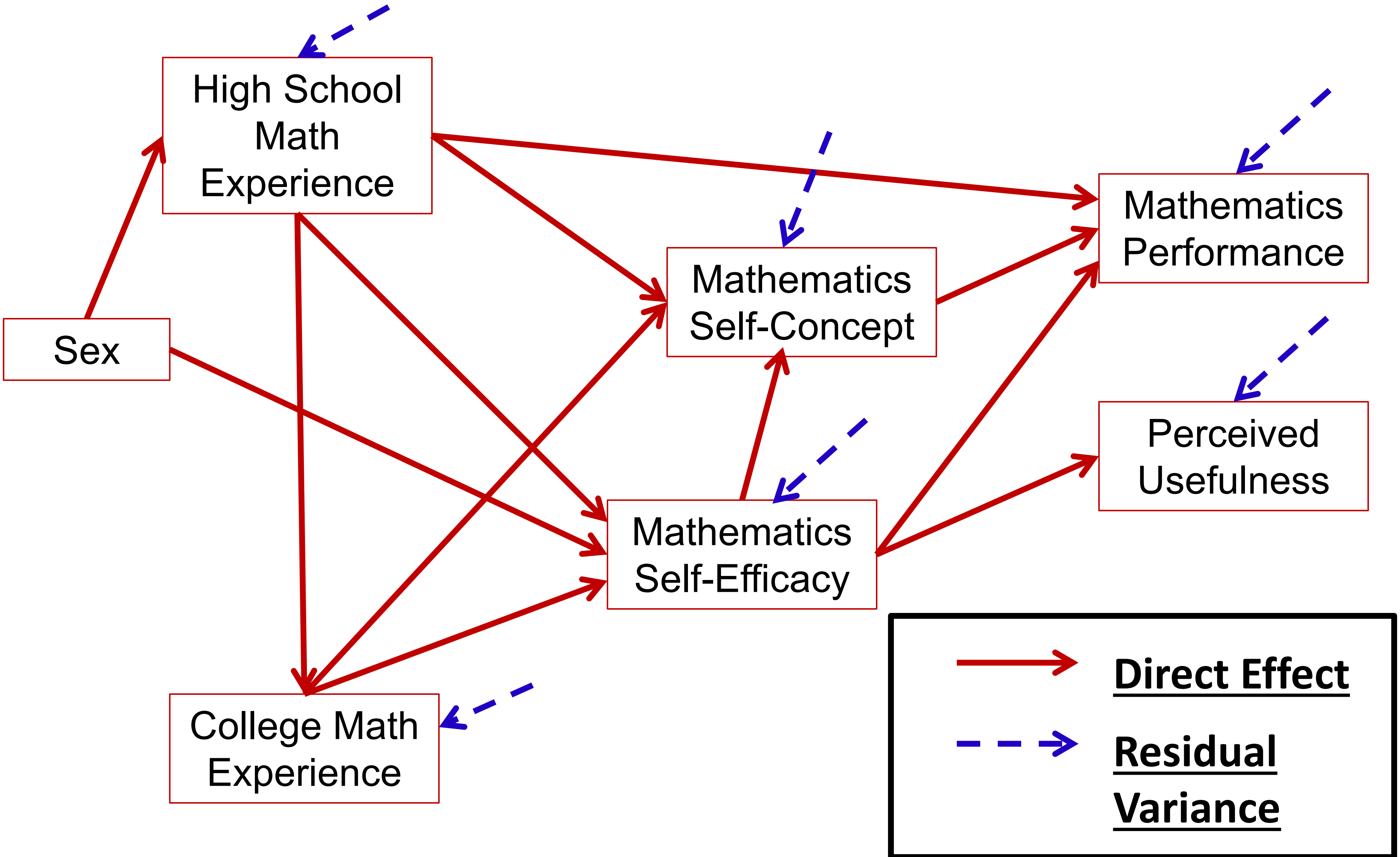
- Sample of 350 undergraduates (229 women, 121 men)
  - In simulation, 10% of variables were missing (using missing completely at random mechanism)

- 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

- Sex (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

- Multivariate linear models are statistical methods that, when using an identity link, assume the variables in an analysis follow a multivariate normal distribution
  - Mean vectors
  - Covariance matrices

- By specifying a set of regression equations that are estimated simultaneously, a very specific covariance matrix is implied

- As with all multivariate models, the key to multivariate linear models is finding an approximation to the unstructured (saturated) covariance matrix
  - With fewer parameters, if possible

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

MULTIVARIATE LINEAR MODELS VIA PATH ANALYSIS SOFTWARE AND PACKAGES

Multivariate Regression

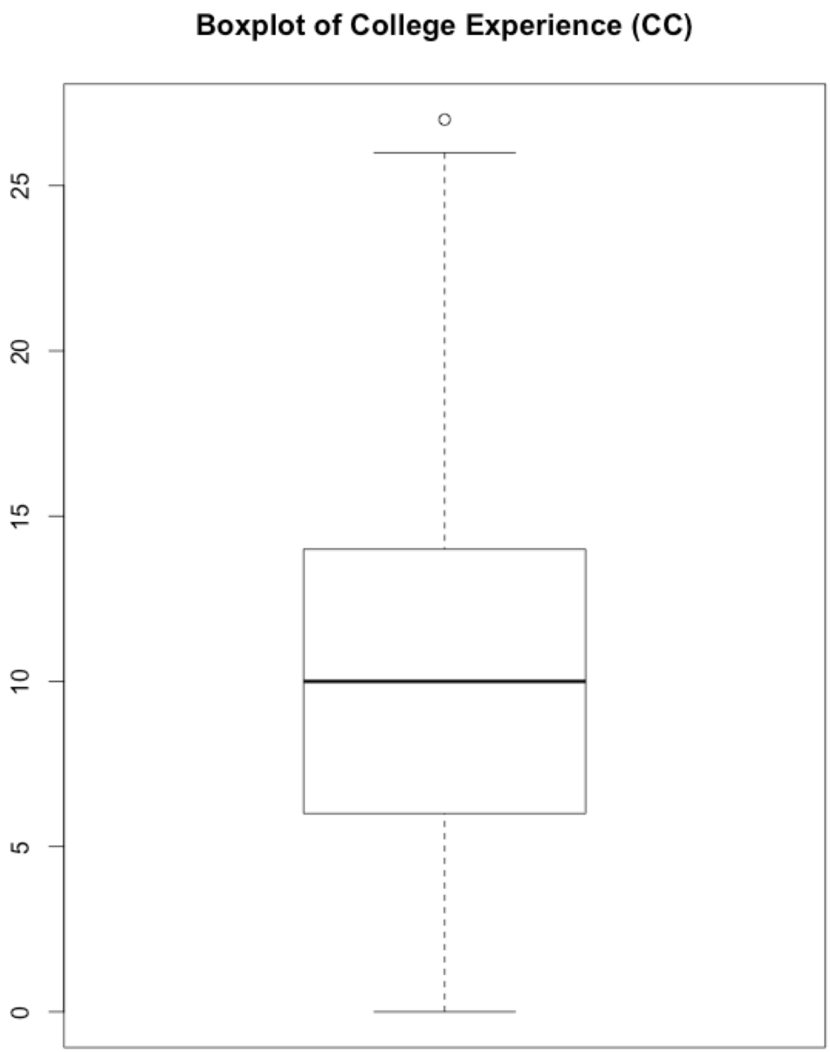
- We begin with a multivariate regression model:
  - Predicting mathematics performance (PERF) with female (F), college math experience (CC), and the interaction between female and college math experience (FxCC)
  - Predicting perceived usefulness (USE) with female (F), college math experience (CC), and the interaction between female and college math experience (FxCC)

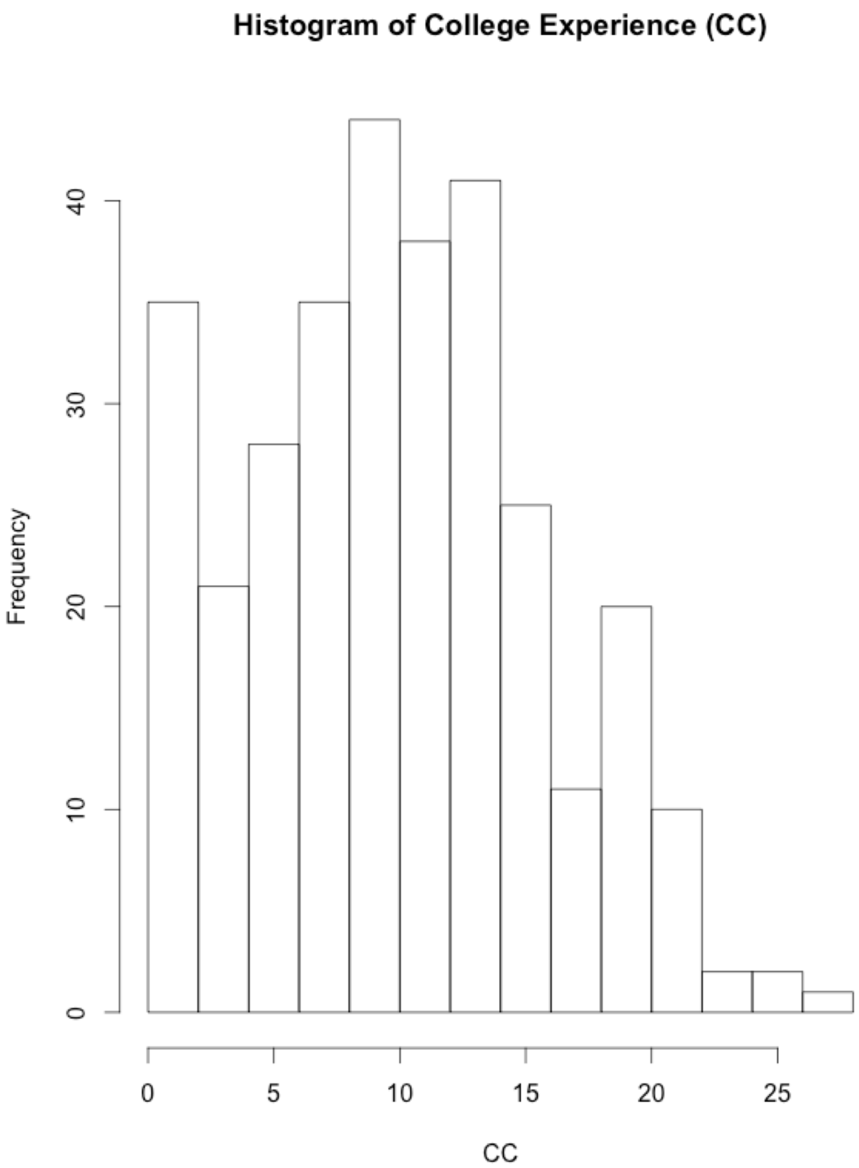
$$PERF_i = \beta_{0,PERF} + \beta_{F,PERF}F_i + \beta_{CC,PERF}CC_i + \beta_{F*CC,PERF}F_iCC_i + e_{i,PERF}$$
$$USE_i = \beta_{0,USE} + \beta_{F,USE}F_i + \beta_{CC,USE}CC_i + \beta_{F*CC,USE}F_iCC_i + e_{i,USE}$$

- 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:

$$\begin{bmatrix} e_{i,PERF} \\ e_{i,USE} \end{bmatrix} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{e,PERF} & \sigma_{e,PERF,USE} \\ \sigma_{e,PERF,USE} & \sigma^2_{e,USE} \end{bmatrix}\right)$$

Before Continuing: We will Center CC at 10





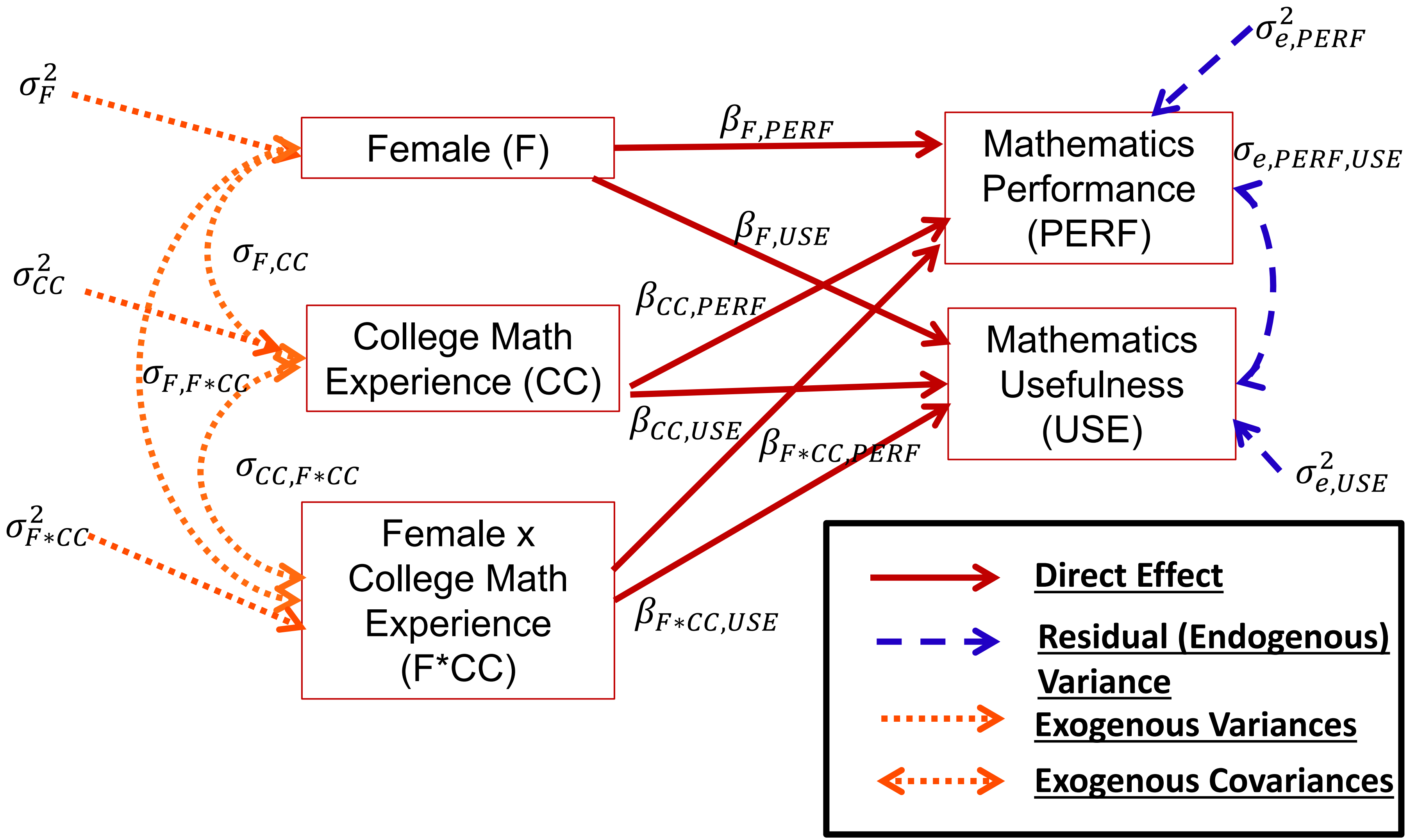
Types of Variables in the Analysis

- 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 our example Mathematics Performance (PERF) and Mathematics Usefulness (USE)
  - In univariate linear regression, the **dependent variable** is the only endogenous variable in an analysis

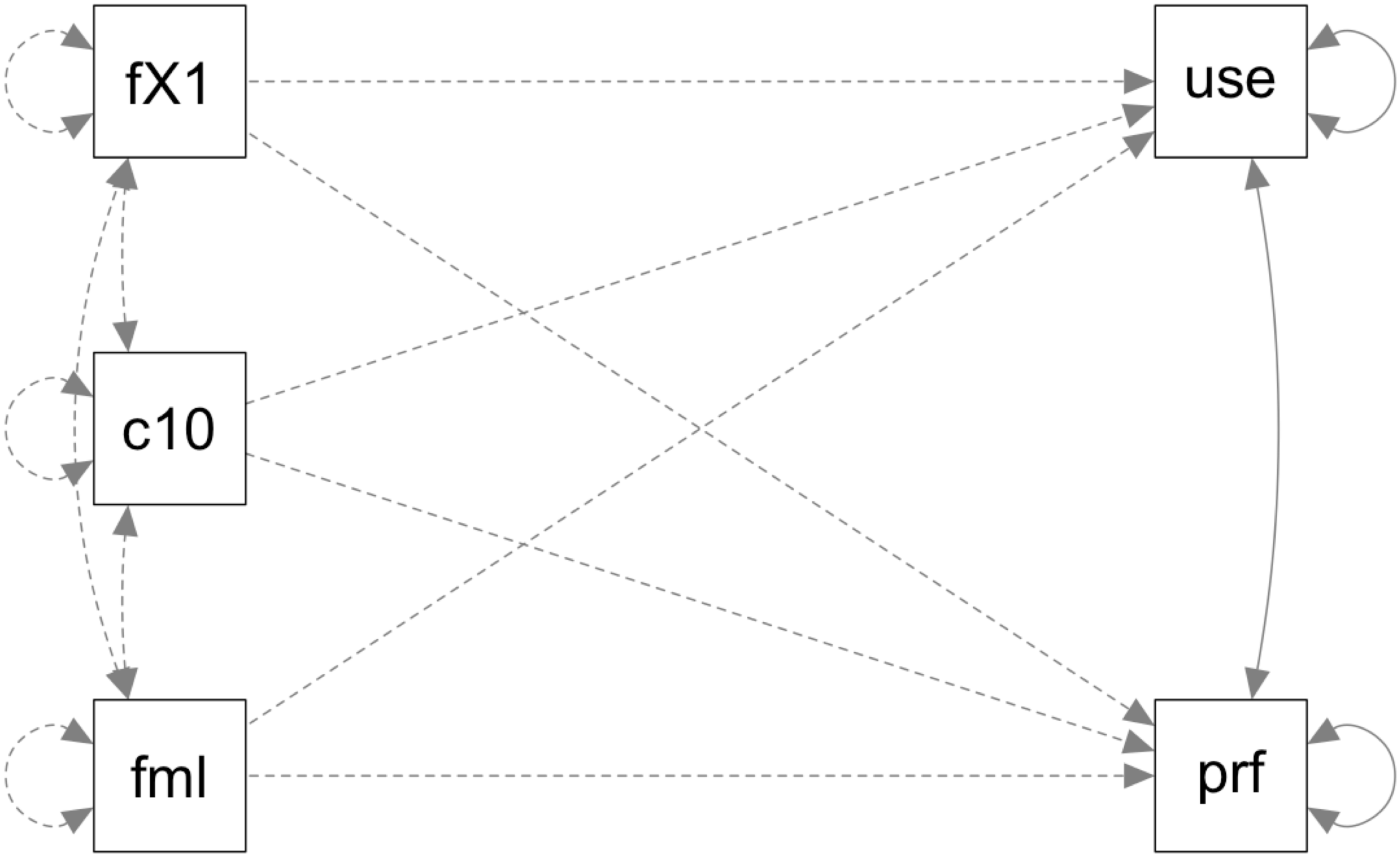
- Exogenous variable(s): variables whose variability ***is not explained*** by any variables in a model
  - In our example Female (F), college experience (CC), and the interaction (FxCC)
  - In linear regression, the **independent variable**(s) are the exogenous variables in the analysis

Multivariate Linear Regression Path Diagram



THE UNIVERSITY OF KANSAS

R's Version of the Path Diagram



Labeling Variables

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

- The exogenous (independent) variables are:
  - Female (F), college experience (CC), and the interaction of Female and college experience (F*CC)

THE UNIVERSITY OF KANSAS

Multivariate Regression in R Using the lavaan Package

```
#Building Analysis Model #1: an empty model-----------------------------------------------------------------
#NOTE: BECAUSE ALL VARIABLES ARE PUT INTO THE LIKELIHOOD FUNCTION, TO DO LIKELIHOOD RATIO TESTS, WE HAVE TO
#      CONSTRUCT THE FULL MODEL BUT MAKE THE REGRESSION COEFFICIENTS EQUAL TO ZERO

#analysis syntax
model01.syntax = "

#Means:
perf ~ 1 + 0*female + 0*cc10 + 0*femXcc10
use  ~ 1 + 0*female + 0*cc10 + 0*femXcc10

#Variances:
perf ~~ perf
use  ~~ use

#Covariance:
perf ~~ use

"

#analysis estimation
model01.fit = sem(model01.syntax, data=data01, conditional.x=TRUE, fixed.x = TRUE, mimic = "MPLUS", estimator = "MLR")
```

By putting 0* in front of each of the variables, we are allowing them to be in the likelihood (for model comparisons) but not predict either DV

- **A note about path analysis software:**
  - Most packages put all variables into the likelihood function (Mplus does not)
  - So, you must start with all variables in the model for LRTs

Multivariate Regression Model Parameters

- Lavaan considers all five variables to be part of a multivariate normal distribution, so the unstructured (saturated) model has a total of 20 parameters:
  - 5 means
  - 5 variances
  - 10 covariances (5-choose-2 or 5*(5-1)/2))

- The model itself has 14 parameters:
  - 5 intercepts
  - 0 regression slopes (but we'll add these next)
  - 2 residual variances
  - 1 residual covariance
  - 3 exogenous variances
  - 3 exogenous covariances

- Lavaan will estimate two models for each analysis: H0 (your model) and H1 (saturated model)

- Degrees of DF in path models come from comparing the saturated model number of parameters with the parameters estimated
  - Parameters available 20 – 14 parameters estimated = 6 df

- Therefore, this model will not fit perfectly – model fit statistics will be available

Output from Lavaan: Summary Statement

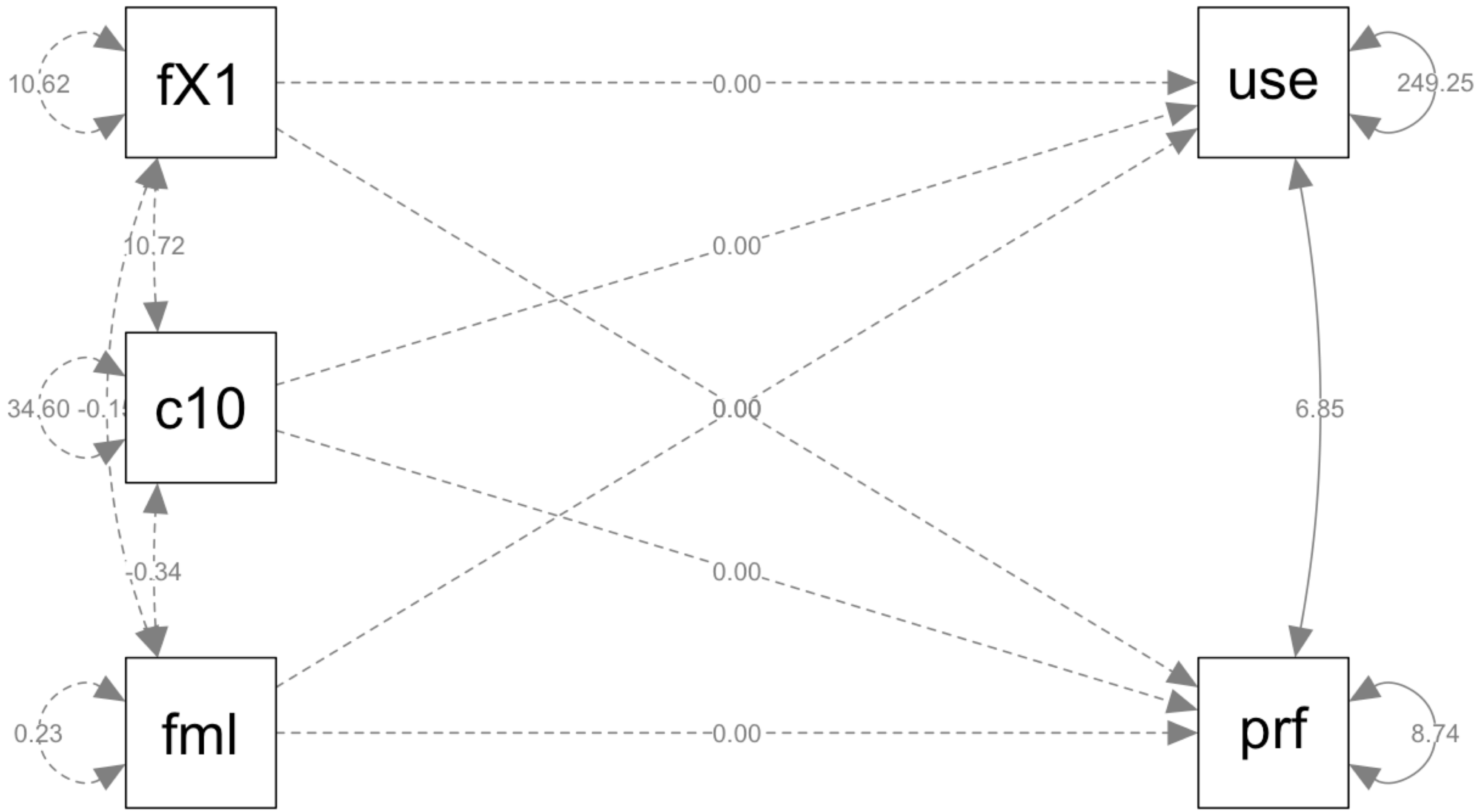
```
lavaan (0.5-20) converged normally after  28 iterations

  Number of observations                           350

  Number of missing patterns                         7

  Estimator                                         ML      Robust
  Minimum Function Test Statistic               22.307      22.204
  Degrees of freedom                                 6           6
  P-value (Chi-square)                           0.001       0.001
  Scaling correction factor                                  1.005
    for the Yuan-Bentler correction (Mplus variant)

Model test baseline model:

  Minimum Function Test Statistic               28.371      27.908
  Degrees of freedom                                 7           7
  P-value                                        0.000       0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.237       0.225
  Tucker-Lewis Index (TLI)                       0.110       0.096

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -4073.253   -4073.253
  Scaling correction factor                                  1.028
    for the MLR correction
  Loglikelihood unrestricted model (H1)      -4062.099   -4062.099
  Scaling correction factor                                  1.015
    for the MLR correction

  Number of free parameters                          5           5
  Akaike (AIC)                                8156.505    8156.505
  Bayesian (BIC)                              8175.795    8175.795
  Sample-size adjusted Bayesian (BIC)         8159.933    8159.933

Root Mean Square Error of Approximation:

  RMSEA                                          0.088       0.088
  90 Percent Confidence Interval          0.051  0.129       0.051  0.128
  P-value RMSEA <= 0.05                          0.046       0.047

Standardized Root Mean Square Residual:

  SRMR                                           0.077       0.077

Parameter Estimates:

  Information                                 Observed
  Standard Errors                   Robust.huber.white
```

```
Regressions:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
  perf ~
    female            0.000                                0.000    0.000
    cc10              0.000                                0.000    0.000
    femXcc10          0.000                                0.000    0.000
  use ~
    female            0.000                                0.000    0.000
    cc10              0.000                                0.000    0.000
    femXcc10          0.000                                0.000    0.000

Covariances:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
  perf ~~
    use               6.847    2.850    2.403    0.016    6.847    0.147

Intercepts:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
    perf             13.959    0.174   80.442    0.000   13.959    4.721
    use              52.440    0.872   60.140    0.000   52.440    3.322

Variances:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
    perf              8.742    0.754   11.596    0.000    8.742    1.000
    use             249.245   19.212   12.973    0.000  249.245    1.000
```

Note:

No information about exogenous variables (from fixed.x=TRUE option)

Path Diagram with Numbers Shown



Output from lavaan: "Fitted" and Saturated Covariance Matrix

```
> fitted(model01.fit)
$cov
         perf    use     female  cc10    fmXc10
perf      8.742
use       6.847 249.245
female    0.000   0.000   0.226
cc10      0.000   0.000  -0.335  34.600
femXcc10  0.000   0.000  -0.146  10.723  10.616

$mean
   perf     use   female     cc10 femXcc10
 13.959  52.440    0.346    0.320   -0.211
```

```
> inspect(model01.fit, what="sampstat.h1")
$cov
         perf    use     female  cc10    fmXc10
perf      8.730
use       6.788 249.254
female    0.070   0.341   0.226
cc10      4.123   8.751  -0.335  34.600
femXcc10  1.920   4.500  -0.146  10.723  10.616

$mean
   perf     use   female     cc10 femXcc10
 13.946  52.468    0.346    0.320   -0.211
```

- The fitted covariance matrix shows you what the model implies the variances and covariances should be
- Here the exogenous variables are provided by sample estimates (fitted.x=TRUE)
- Model parameters provide the endogenous parameters
- The lower matrix is the saturated model matrix

Output from lavaan: Residual Covariance Matrices

```
> residuals(model01.fit, type = "raw")
$type
[1] "raw"

$cov
        perf   use    female cc10   fmXc10
perf    -0.012
use     -0.059  0.009
female   0.070  0.341  0.000
cc10     4.123  8.751  0.000  0.000
femXcc10 1.920  4.500  0.000  0.000  0.000

$mean
   perf      use   female     cc10 femXcc10
 -0.013    0.029    0.000    0.000    0.000
```

- The "raw" residuals are the difference between the model implied covariance matrix and the H1 (saturated model) covariance matrix/mean vector

METHODS OF EXAMINING MODEL FIT

Methods of Model Fit

- Model-data fit is of utmost concern when building models with multivariate outcomes
- If a model does not fit the data:
  - Parameter estimates may be biased
  - Standard errors of estimates may be biased
  - Inferences made from the model may be wrong
  - If the saturated model fit is wrong, then the LRTs will be inaccurate
- Examining model fit is the first step in multivariate models
- That said, not all “good-fitting” models are useful…
  - …model fit just allows you to talk about your model…there may be nothing of significance (statistically or practically) in your results, though

Types of Model Fit Information

- Model fit information for models where outcomes are conditionally MVN* come in several types, but all are based on the premise that any model mean and covariance structure must fit as well as the saturated mean vector and covariance matrix model

  ***If model outcomes are not conditionally MVN, model fit is very different**

- All possible models/structures **are nested within** the saturated mean vector and covariance matrix model
  - Most model fit statistics come from comparing any model/structure with the saturated model

- Indices shown first are called "global" model fit indices
  - Report fit of model globally (as opposed to locally for specific parameters)

Example lavaan Model Fit Output

```
lavaan (0.5-20) converged normally after  28 iterations

  Number of observations                           350

  Number of missing patterns                         7

  Estimator                                         ML      Robust
  Minimum Function Test Statistic               22.307      22.204
  Degrees of freedom                                 6           6
  P-value (Chi-square)                           0.001       0.001
  Scaling correction factor                                  1.005
    for the Yuan-Bentler correction (Mplus variant)

Model test baseline model:

  Minimum Function Test Statistic               28.371      27.908
  Degrees of freedom                                 7           7
  P-value                                        0.000       0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.237       0.225
  Tucker-Lewis Index (TLI)                       0.110       0.096

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -4073.253   -4073.253
  Scaling correction factor                                  1.028
    for the MLR correction
  Loglikelihood unrestricted model (H1)      -4062.099   -4062.099
  Scaling correction factor                                  1.015
    for the MLR correction

  Number of free parameters                          5           5
  Akaike (AIC)                                8156.505    8156.505
  Bayesian (BIC)                              8175.795    8175.795
  Sample-size adjusted Bayesian (BIC)         8159.933    8159.933

Root Mean Square Error of Approximation:

  RMSEA                                          0.088       0.088
  90 Percent Confidence Interval          0.051  0.129       0.051  0.128
  P-value RMSEA <= 0.05                          0.046       0.047

Standardized Root Mean Square Residual:

  SRMR                                           0.077       0.077

Parameter Estimates:

  Information                                 Observed
  Standard Errors                   Robust.huber.white
```

The fit.measures=TRUE Model Fit Statistics

- **Unlabeled section**
  - **Likelihood ratio test versus the saturated model**
  - **Testing if your model fits as well as the saturated model**
- Model test baseline model
  - Likelihood ratio test pitting the saturated model against the independent variables model
  - Testing whether any variables have non-zero covariances (significant correlations)
- User model versus baseline model
  - CFI
  - TLI
- Loglikelihood and Information Criteria
  - Likelihood ratio tests (nested models)
  - Information criteria comparisons (non-nested models)
- Root Mean Square Error of Approximation
  - How far off a model is from the saturated model, per degree of freedom
- Standardized Root Mean Square Residual
  - How far off a model's correlations are from the saturated model correlations

Indices of Global Model Fit

- Primary: obtained model **χ^2** (from Model test baseline model) – here we use the MLR rescaled **χ^2** from the "Robust" Column
  - χ^2 is evaluated based on model df (difference in parameters between your CFA model and the saturated model)
  - Tests null hypothesis that ***this*** model (H_0) fits equally to ***saturated model*** (H_1) so significance is undesirable (smaller χ^2, bigger p-value is better)
    - Means saturated model is estimated **automatically** for each model analyzed
  - Just using χ^2 is insufficient, however:
    - Distribution doesn't behave like a true χ^2 if sample sizes are small (or, if not using MLR, if items are non-normally distributed)
    - Obtained χ^2 depends largely on sample size
    - Some mention this is an unreasonable null hypothesis (perfect fit??)

- Because of these issues, alternative measures of fit are usually used in conjunction with the χ^2 test of model fit
  - Absolute Fit Indices (besides χ^2)
  - Parsimony-Corrected; Comparative (Incremental) Fit Indices

Chi-Square Test of Model Fit

- The Chi-Square Test of Model Fit provides a likelihood ratio test comparing the current model to the **saturated (unstructured) model**:
  - The value is -2 times the difference in log-likelihoods (rescaled if MLR)
  - The degrees of freedom is the difference in the number of estimated model parameters
  - The p-value is from the Chi-square distribution
- **If this test has a significant p-value:**
  - The current model (H_0) is rejected – the model fit is significantly worse than the full model
  - In latent variable models, this test is usually ignored
    - Said to be overly sensitive
- **If this test does not have a significant p-value:**
  - The current model (H_0) is not rejected – **fits equivalently to full model**

Where the Saturated Model Test Comes From

- The saturated model LRT comes from a likelihood ratio test of the current model with the saturated model
- If using MLR (Robust method), then this LRT is rescaled based on the estimated scaling factors of both models
- This same information can be obtained from:
  - Loglikelihood model output section
  - anova() function comparing fit for current and saturated models

THE UNIVERSITY OF KANSAS

Calculating the LRT for Global Fit Test for Model 04

- From the lavaan output:

```
Estimator                                         ML       Robust
Minimum Function Test Statistic               22.307       22.204
Degrees of freedom                                 6            6
P-value (Chi-square)                           0.001        0.001
Scaling correction factor                                   1.005
  for the Yuan-Bentler correction (Mplus variant)
```

```
Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -4073.253   -4073.253
  Scaling correction factor                                  1.028
    for the MLR correction
  Loglikelihood unrestricted model (H1)      -4062.099   -4062.099
  Scaling correction factor                                  1.015
    for the MLR correction
```

- Calculation:
  - 14 parameters in our model; 20 in saturated model
  - Scaling correction factor:

$$c_{LR} = \left| \frac{(q_{restricted})(c_{restricted}) - (q_{full})(c_{full})}{(q_{restricted} - q_{full})} \right| = 1.005$$

  - $\chi^2 = \frac{22.307}{1.005} = 22.204$
  - DF = 6

- Conclusion: this model fit significantly worse than the saturated model
  - And it should—especially if any of our predictors have non-zero betas

Saturated Model LRT and Loglikelihood Output

```
Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)               -4073.253   -4073.253
  Scaling correction factor                                   1.028
    for the MLR correction
  Loglikelihood unrestricted model (H1)       -4062.099   -4062.099
  Scaling correction factor                                   1.015
    for the MLR correction
```

- If the loglikelihoods of the current model ("User model" or H_0) are equal to the loglikelihoods of the saturated model ("Unrestriced model" or H_1), then you are running a model that is equivalent to the saturated model
  - **<u>No other model fit will be available or useful</u>**

THE UNIVERSITY OF KANSAS

The fit.measures=TRUE Model Fit Statistics

- ~~Unlabeled section~~
  - ~~Likelihood ratio test versus the saturated model~~
  - ~~Testing if your model fits as well as the saturated model~~

- **Model test baseline model**
  - **Likelihood ratio test pitting the saturated model against the independent variables model**
  - **Testing whether any variables have non-zero covariances (significant correlations)**

- User model versus baseline model
  - CFI
  - TLI

- Loglikelihood and Information Criteria
  - Likelihood ratio tests (nested models)
  - Information criteria comparisons (non-nested models)

- Root Mean Square Error of Approximation
  - How far off a model is from the saturated model, per degree of freedom

- Standardized Root Mean Square Residual
  - How far off a model's correlations are from the saturated model correlations

Model Test Baseline Model

- The "model test baseline model" section provides a LRT:
  - Comparing the saturated (unstructured) model with an independent variables model (called the baseline model)

```
Model test baseline model:

  Minimum Function Test Statistic               28.371      27.908
  Degrees of freedom                                 7           7
  P-value                                        0.000       0.000
```

- Here, the "null" model is the baseline (the independent variables model)
  - If the test is significant, this means that at least one (and likely more than one) variable has a significant covariance (and correlation)
  - If the test is not significant, this means that the independence model is appropriate
    - This is not likely to happen
    - But if it does, there are virtually no other models that will be significant
- Not often reported as it is likely variables are correlated

The fit.measures=TRUE Model Fit Statistics

- ~~Unlabeled section~~
  - ~~Likelihood ratio test versus the saturated model~~
  - ~~Testing if your model fits as well as the saturated model~~

- ~~Model test baseline model~~
  - ~~Likelihood ratio test pitting the saturated model against the independent variables model~~
  - ~~Testing whether any variables have non-zero covariances (significant correlations)~~

- **User model versus baseline model**
  - **CFI**
  - **TLI**

- Loglikelihood and Information Criteria
  - Likelihood ratio tests (nested models)
  - Information criteria comparisons (non-nested models)

- Root Mean Square Error of Approximation
  - How far off a model is from the saturated model, per degree of freedom

- Standardized Root Mean Square Residual
  - How far off a model's correlations are from the saturated model correlations

User Model Versus Baseline Model Section

- The "User model versus baseline model" section provides two additional measures of model fit comparing the current (user) model to the baseline (independent variables) model

```
User model versus baseline model:

  Comparative Fit Index (CFI)                    0.237       0.225
  Tucker-Lewis Index (TLI)                       0.110       0.096
```

- CFI stands for Comparative Fit Index
  - Higher is better (above .95 indicates good fit)
- TLI stands for Tucker Lewis Index
  - Higher is better (above .95 indicates good fit)

Comparative (Incremental) Fit Indices

- Fit evaluated relative to a 'null' model (of 0 covariances)
  - Relative to that, your model should be great!

T = target (current/estimated) model
N = null (baseline/independent variables) model

- **CFI: Comparative Fit Index**
  - Based on idea of the chi-square non-centrality parameter: (χ^2 – df)
  - $CFI = 1 - \frac{\max(\chi_T^2 - df_T, 0)}{\max(\chi_T^2 - df_T, \chi_N^2 - df_N, 0)}$
  - From 0 to 1: bigger is better, > .90 = "acceptable", > .95 = "good"

- **TLI: Tucker-Lewis Index** (= **Non-Normed Fit Index**)
  - $TLI = \frac{\frac{\chi_N^2}{df_N} - \frac{\chi_T^2}{df_T}}{\frac{\chi_N^2}{df_N} - 1}$
  - From <0 to >1, bigger is better, >.95 = "good"

Information Criteria Output

- The information criteria output provides relative fit statistics:

```
Number of free parameters                          5          5
Akaike (AIC)                                8156.505   8156.505
Bayesian (BIC)                              8175.795   8175.795
Sample-size adjusted Bayesian (BIC)         8159.933   8159.933
```

  - AIC: Akaike Information Criterion
  - BIC: Bayesian Information Criterion (also called Schwarz's criterion)
  - Sample-size Adjusted BIC

- These statistics weight the information given by the parameter values by the parsimony of the model (the number of model parameters)
  - For all statistics, the smaller number is better
- The core of these statistics is -2*log-likelihood

The fit.measures=TRUE Model Fit Statistics

- ~~Unlabeled section~~
  - ~~Likelihood ratio test versus the saturated model~~
  - ~~Testing if your model fits as well as the saturated model~~

- ~~Model test baseline model~~
  - ~~Likelihood ratio test pitting the saturated model against the independent variables model~~
  - ~~Testing whether any variables have non-zero covariances (significant correlations)~~

- ~~User model versus baseline model~~
  - ~~CFI~~
  - ~~TLI~~

- **Loglikelihood and Information Criteria**
  - **Likelihood ratio tests (nested models)**
  - **Information criteria comparisons (non-nested models)**

- Root Mean Square Error of Approximation
  - How far off a model is from the saturated model, per degree of freedom

- Standardized Root Mean Square Residual
  - How far off a model's correlations are from the saturated model correlations

Comparing Information Criteria

- Information criteria are relative tests of fit

Number of free parameters	5	5
Akaike (AIC)	8156.505	8156.505
Bayesian (BIC)	8175.795	8175.795
Sample-size adjusted Bayesian (BIC)	8159.933	8159.933

- The are calculated based on the log-likelihood of the model, factoring in a penalty for number of parameters (plus other things)

- They should never be used to compare nested models
  - The likelihood ratio test is the most powerful test statistic to use for nested models

- When comparing non-nested models, first choose a statistic
  - AIC, BIC, or Sample-size Adjusted BIC are what are given by default

- The preferred model is the one with the lowest value of that statistic

The fit.measures=TRUE Model Fit Statistics

- ~~Unlabeled section~~
  - ~~Likelihood ratio test versus the saturated model~~
  - ~~Testing if your model fits as well as the saturated model~~
- ~~Model test baseline model~~
  - ~~Likelihood ratio test pitting the saturated model against the independent variables model~~
  - ~~Testing whether any variables have non-zero covariances (significant correlations)~~
- ~~User model versus baseline model~~
  - ~~CFI~~
  - ~~TLI~~
- ~~Loglikelihood and Information Criteria~~
  - ~~Likelihood ratio tests (nested models)~~
  - ~~Information criteria comparisons (non-nested models)~~
- **Root Mean Square Error of Approximation**
  - **How far off a model is from the saturated model, per degree of freedom**
- Standardized Root Mean Square Residual
  - How far off a model's correlations are from the saturated model correlations

THE UNIVERSITY OF KANSAS

Indices of Global Model Fit

Parsimony-Corrected: **RMSEA**

- **Root Mean Square Error of Approximation**
- Uses comparison with CFA model and saturated model
  - χ^2 listed here from first part of lavaan output
- Relies on a non-centrality parameter (NCP)
  - Indexes how far off your model is → χ^2 distribution shoved over
  - NCP → $d = (\chi^2 - df) / (N-1)$ Then, RMSEA = SQRT(d/df)
    - df is difference between # parameters in CFA model and saturated model
  - RMSEA ranges from 0 to 1; smaller is better
    - < .05 or .06 = "good", .05 to .08 = "acceptable",
.08 to .10 = "mediocre", and >.10 = "unacceptable"
  - In addition to point estimate, get 90% confidence interval
  - RMSEA penalizes for model complexity – it's discrepancy in fit per df left in model (but not sensitive to N, although CI can be)
  - Test of "close fit": null hypothesis that RMSEA ≤ .05

RMSEA (Root Mean Square Error of Approximation)

- The RMSEA is an index of model fit where 0 indicates perfect fit (smaller is better):

```
Root Mean Square Error of Approximation:

  RMSEA                                          0.088        0.088
  90 Percent Confidence Interval         0.051  0.129        0.051  0.128
  P-value RMSEA <= 0.05                          0.046        0.047
```

- RMSEA is based on the approximated covariance matrix

- The goal is a model with an RMSEA less than .05
  - Although there is some flexibility

- The result above indicates our model fits poorly (RMSEA of .0088)

The fit.measures=TRUE Model Fit Statistics

- ~~Unlabeled section~~
  - ~~Likelihood ratio test versus the saturated model~~
  - ~~Testing if your model fits as well as the saturated model~~

- ~~Model test baseline model~~
  - ~~Likelihood ratio test pitting the saturated model against the independent variables model~~
  - ~~Testing whether any variables have non-zero covariances (significant correlations)~~

- ~~User model versus baseline model~~
  - ~~CFI~~
  - ~~TLI~~

- ~~Loglikelihood and Information Criteria~~
  - ~~Likelihood ratio tests (nested models)~~
  - ~~Information criteria comparisons (non-nested models)~~

- ~~Root Mean Square Error of Approximation~~
  - ~~How far off a model is from the saturated model, per degree of freedom~~

- **Standardized Root Mean Square Residual**
  - **How far off a model's correlations are from the saturated model correlations**

THE UNIVERSITY OF KANSAS

Standardized Root Mean Squared Residual

- The SRMR (standardized root mean square residual) provides the average standardized difference between:
  - The estimated covariance matrix of the saturated model
  - The estimated covariance matrix of the current model

```
Standardized Root Mean Square Residual:

  SRMR                                    0.077       0.077
```

- Lower is better (some suggest less than 0.08)

LOCAL MODEL FIT MEASURES

“Local” Measures of Model (Mis)Fit

- Local measures of model (mis)fit are statistics that point to the location (typically of a covariance matrix) where a model may not fit well
  - As opposed to “global” measures that indicate a model fit overall

- Local measures of model (mis)fit are typically of two types:
  - Residual covariance matrices (unstandardized, standardized, or normalized)
    - The difference between the model’s estimated covariance matrix and the saturated model’s estimated covariance matrix
    - These were used for the SRMR
  - Model “modification indices”
    - 1-degree of freedom hypothesis tests for the improvement of the model LRT if one more parameter was allowed to be estimated

Residual Covariance Matrices

- Residual covariance matrices are used to figure out how to best improve model misfit

- The "raw" or "unstandardized" residual covariance matrix for the model literally takes the difference between model implied and saturated model covariance matrices

- I often prefer "normalized" versions of these matrices
  - We can inspect the normalized residual covariance matrix (like z-scores) to see where our biggest misfit occurs

```
> residuals(model01.fit, type = "normalized")
$type
[1] "normalized"

$cov
        perf   use    female cc10   fmXc10
perf    -0.016
use     -0.021  0.000
female   0.848  0.856  0.000
cc10     3.916  1.591  0.000  0.000
femXcc10 2.775  1.636  0.000  0.000  0.000

$mean
    perf      use   female     cc10 femXcc10
  -0.078    0.033    0.000    0.000    0.000
```


THE UNIVERSITY OF KANSAS

Modification Indices: More Help for Fit

- As we used Maximum Likelihood to estimate our model, another useful feature is that of the modification indices
  - Modification indices, also called Score or LaGrangian Multiplier tests, attempt to suggest the change in the log-likelihood for adding a given model parameter (larger values indicate a better fit for adding the parameter)

```
> modindices(model01.fit)
   lhs op     rhs     mi mi.scaled   epc sepc.lv sepc.all sepc.nox
2 perf  ~  female  0.811     0.808 0.326   0.326    0.052    0.110
3 perf  ~    cc10 15.420    15.348 0.121   0.121    0.240    0.041
4 perf  ~ femXcc10  9.285     9.242 0.169   0.169    0.187    0.057
6  use  ~  female  0.436     0.434 1.204   1.204    0.036    0.076
7  use  ~    cc10  1.134     1.128 0.166   0.166    0.062    0.010
8  use  ~ femXcc10  1.196     1.190 0.307   0.307    0.063    0.019
```

- mi column: the expected value of the LRT of the current model and a model where this parameter was added
- mi.scaled column: the scaled (robust) LRT
  - Should be bigger than 3.84 for 1 df
  - Practice is to find values that are much higher (say 10 or more)
- epc column: expected value of the parameter in the model where this parameter was added

ADDING PREDICTORS TO THE MODEL

Adding Predictors: Removing Zero Values from Parameters

```
#model 02: all parameters included ----------------------------------------------------------------------
model02.syntax = "

#Means:
perf ~ 1 + (p_f)*female + (p_cc)*cc10 + (p_f_cc)*femXcc10
use  ~ 1 + (u_f)*female + (u_cc)*cc10 + (u_f_cc)*femXcc10

#Variances:
perf ~~ perf
use  ~~ use

#Covariance:
perf ~~ use

#Defined parameters (glht() analog in lavaan)
cc_perf_fem := p_cc + p_f_cc
cc_use_fem := u_cc + u_f_cc

"
#analysis estimation
model02.fit = sem(model02.syntax, data=data01, conditional.x=TRUE, fixed.x = TRUE, mimic = "MPLUS", estimator = "MLR")
```

First Question: Which Model "Fits" Better?

- After adding the predictors (estimating their betas) to the model, we must first ask which model fits better
- A likelihood ratio test (LRT) can be performed comparing model02 (with predictors) and model01 (without)
- **Which model is the null model?**
- **Which model is the alternative model?**
- **What is the null hypothesis?**
- **What is the alterative hypothesis?**

LRT With Scaled Chi-Squares

- R makes the scaled Chi-square LRT easy…use the anova() function and it will rescale the Chi-squares automatically

```
> anova(model01.fit, model02.fit)
Scaled Chi Square Difference Test (method = "satorra.bentler.2001")

            Df    AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
model02.fit  0 8146.2 8188.6  0.000
model01.fit  6 8156.5 8175.8 22.307     22.204       6   0.001112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Here we see that we reject model01 (the null model)
- So we conclude that **at least** one beta value was significantly different from zero

Step 2: Inspect Model Fit

- Next we inspect the model fit of model02:

```
> summary(model02.fit, standardized=TRUE, fit.measures=TRUE)
lavaan (0.5-20) converged normally after  57 iterations

  Number of observations                           350

  Number of missing patterns                         7

  Estimator                                         ML      Robust
  Minimum Function Test Statistic                0.000       0.000
  Degrees of freedom                                 0           0
  Scaling correction factor                                     NA
    for the Yuan-Bentler correction (Mplus variant)

Model test baseline model:

  Minimum Function Test Statistic               28.371      27.908
  Degrees of freedom                                 7           7
  P-value                                        0.000       0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    1.000       1.000
  Tucker-Lewis Index (TLI)                       1.000       1.000

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -4062.099   -4062.099
  Loglikelihood unrestricted model (H1)      -4062.099   -4062.099

  Number of free parameters                         11          11
  Akaike (AIC)                                8146.198    8146.198
  Bayesian (BIC)                              8188.635    8188.635
  Sample-size adjusted Bayesian (BIC)         8153.739    8153.739

Root Mean Square Error of Approximation:

  RMSEA                                          0.000       0.000
  90 Percent Confidence Interval          0.000  0.000       0.000  0.000
  P-value RMSEA <= 0.05                          1.000       1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.000       0.000

Parameter Estimates:

  Information                                 Observed
  Standard Errors                   Robust.huber.white
```

- Model02 has the same log-likelihood as the saturated model…so it is equivalent to the saturated model
  - Therefore it fits perfectly!

- Any path model where **all** exogenous variables predict **all** endogenous variables **AND** all covariances between endogenous variables are estimated is the saturated model

Up Next: Inspect Parameters and Make Interpretations

```
Regressions:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
  perf ~
    female    (p_f)   0.510    0.352    1.448    0.148    0.510    0.082
    cc10     (p_cc)   0.096    0.033    2.931    0.003    0.096    0.191
    fmXcc10 (p_f_)    0.091    0.068    1.341    0.180    0.091    0.100
  use ~
    female    (u_f)   1.960    1.776    1.104    0.270    1.960    0.059
    cc10     (u_cc)   0.192    0.200    0.961    0.337    0.192    0.072
    fmXcc10 (u_f_)    0.257    0.329    0.780    0.436    0.257    0.053

Covariances:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
  perf ~~
    use               5.365    2.794    1.920    0.055    5.365    0.120

Intercepts:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
    perf             13.758    0.209   65.944    0.000   13.758    4.656
    use              51.783    1.128   45.921    0.000   51.783    3.280

Variances:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
    perf              8.124    0.712   11.411    0.000    8.124    0.931
    use             245.747   18.726   13.123    0.000  245.747    0.986

Defined Parameters:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
    cc_perf_fem       0.187    0.059    3.150    0.002    0.187    0.291
    cc_use_fem        0.449    0.261    1.720    0.086    0.449    0.125
```

New Terms: Standardized Parameters

- Standardized parameters are parameters that are transformed by dividing by one or more standard deviations

- Big-picture example: Recall the covariance to correlation formula

$$\text{Correlation}(X, Y) = \frac{\text{Covariance}(X, Y)}{\text{SD}(X) * \text{SD}(Y)}$$

- The correlation is a standardized covariance
- Standardized = units removed

Standardized Regression Parameters

- The standardized regression parameters are similar
- Take the original equation for a simple linear (one predictor) regression:

$$\beta_Y^X = \rho_{X,Y} \frac{\sigma_y}{\sigma_x}$$

  - β_Y^X is interpreted as the increase in units of Y per units of X

- To standardize (std.all in lavaan), remove units:

$$b_Y^X = \beta_Y^X \left(\frac{\sigma_x}{\sigma_y}\right) = \rho_{X,Y}$$

  - b_Y^X is interpreted as the increase in SDs of Y per SDs of X

- **Standardized parameters are useful for comparing effects on different scales**

Questions to Answer about this Model

- What is the effect of college experience on usefulness for males?
- What is the effect of college experience on usefulness for females?
- What is the difference between males and females ratings of usefulness when college experience = 10?
- How did the difference between males and females ratings change for each additional hour of college experience?

Questions to Answer about this Model

- What is the effect of college experience on performance for males?
- What is the effect of college experience on performance for females?
- What is the difference between males and females performance when college experience = 10?
- How did the difference between males and females performance change for each additional hour of college experience?

WRAPPING UP

Multivariate Linear Models with Predictors

- In this lecture we discussed the basics of multivariate linear models with predictors
  - Model specification/identification
  - Model estimation
  - Model fit (necessary, but not sufficient)
  - Model modification and re-estimation
  - Final model parameter interpretation

- There is a lot to the analysis – but what is important to remember is the over-arching principal of multivariate analyses: covariance between variables is important
  - Path models imply very specific covariance structures
  - The validity of the results hinge upon accurately finding an approximation to the covariance matrix