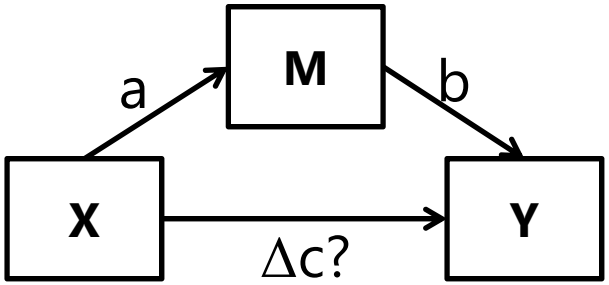
Introduction to Path Analysis and Mediation

- Today's Class:
  - Multivariate models via path analysis
  - Model identification and absolute model fit
  - Mediation and indirect effects

Uses of "Multivariate" Models:

- **(Lectures 5, 6, 7) When y_i is still a single outcome, but:**
  - You have more than one outcome per person created by multiple conditions (e.g., longitudinal or repeated measures designs)
  - When your y_i comes from people nested or clustered in groups, such that you really have multivariate outcomes of a group (e.g., children nested in teachers, people nested in families)
- **When your hypotheses involve more than one y_i:**
  - To compare predictor effect sizes across outcomes (e.g., is a treatment effect bigger on outcome A than outcome B?)
  - You want to test indirect effects among them (i.e., mediation):



In this "path model" M is an outcome of X and a predictor of Y

Indirect effect = Δc? = a*b=0?

Path Models: Pictures and Equations

- Path model: Multivariate models for predicting 2+ outcomes simultaneously for the same unit of analysis
- Most often expressed as a diagram using these conventions:
  - Boxes = observed variables; circles = latent variables (in SEM) or residual
  - One-headed arrow = regression (arrow points from predictor to outcome)
  - Two-headed arrow = residual covariance; intercepts typically not shown

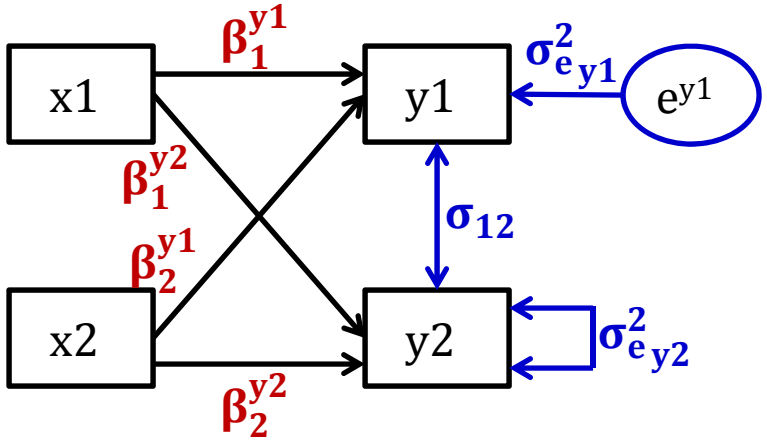
Diagram translates into these simultaneous regression models (in which superscripts denote the outcome of each parameter):

$$y1_i = \beta_0^{y1} + \beta_1^{y1}(x1_i) + \beta_2^{y1}(x2_i) + e_i^{y1}$$
$$y2_i = \beta_0^{y2} + \beta_1^{y2}(x1_i) + \beta_2^{y2}(x2_i) + e_i^{y2}$$

Unstructured R matrix for outcome variances and covariance(s):

$$\begin{bmatrix} \sigma^2_{e_{y1}} & \sigma_{12} \\ \sigma_{12} & \sigma^2_{e_{y2}} \end{bmatrix}$$

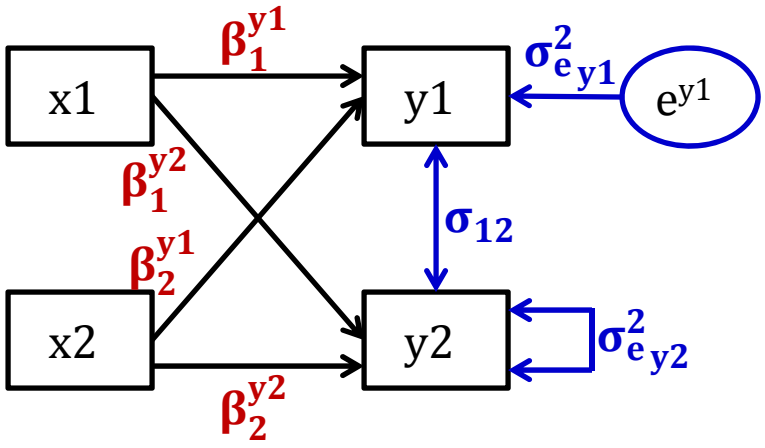
The idea of residual variance is either expressed using a separate circle (as for Y1) or a two-headed arrow into itself (as for Y2).



Multivariate Regression via Path Models

- This example is really just two univariate regression models estimated simultaneously
  - β_1 and β_2 provide the unique effects of x1 and x2 for y1 and y2 outcomes
  - Can calculate R^2 for each outcome

- So why bother to do it this way?
  - To test differences in effect size (e.g., does $\beta_1^{y1} = \beta_2^{y1}$?)
  - To test mediation and indirect effects, in which a variable is both a predictor and an outcome in the same analysis (stay tuned)



If these variables came from a dyad of two persons (1 and 2), this could be an example of an "actor–partner model"

- Arrows within same person = "actor effects"
- Arrows across different people = "partner effects"

2 Types of Path Model Solutions

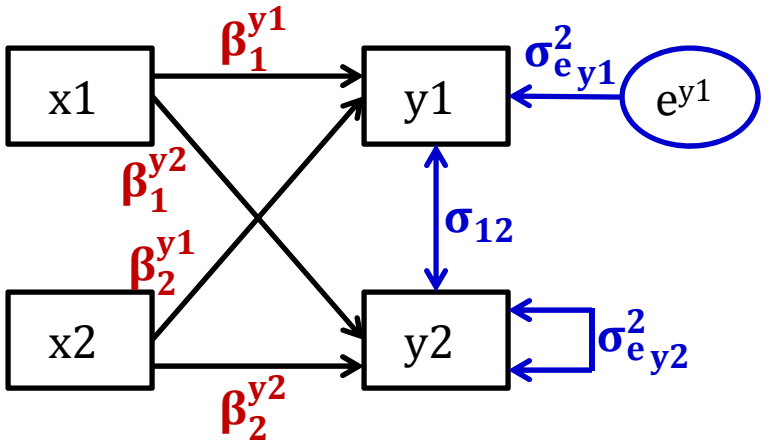
- Unstandardized → predicts scale-sensitive original variables:
  - **Regression Model:** $y1_i = \boldsymbol{\beta_0^{y1}} + \boldsymbol{\beta_1^{y1}}(x1_i) + \boldsymbol{\beta_2^{y1}}(x2_i) + \boldsymbol{e_i^{y1}}$
  - Useful for comparing across groups (whenever absolute values matter)
  - Model parameters predict the intercepts and covariance matrix
  - Variance of **y1** = **[variance explained by model for the means]** + $\boldsymbol{\sigma^2_{e_{y1}}}$

- Standardized → Solution using z-scored versions of variables:
  - Useful when comparing effects within a solution (are then on same scale)
  - Standardized model parameters predict the **variable correlation matrix**
  - Standardized slope = $[\boldsymbol{\beta_1^{y1}} * \mathbf{SD(x1)}] / \mathbf{SD(y1)}$ = **unique correlation**
  - $\mathbf{R^2}$ for **y1** = $\mathbf{1 - standardized\ \sigma^2_{e_{y1}}}$

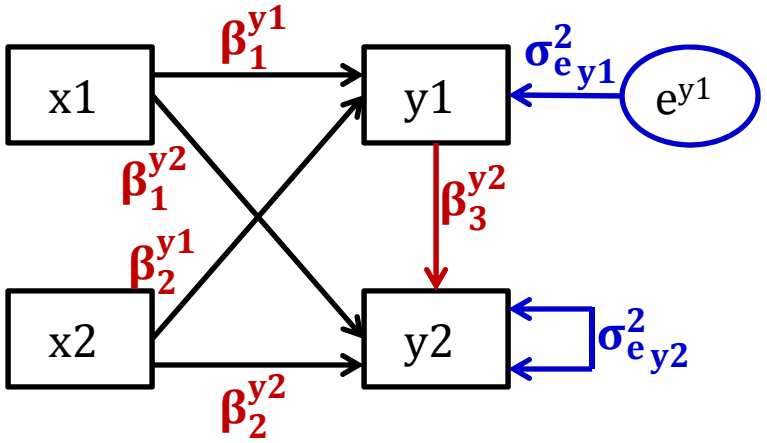
New (and Confusing) Terminology

- Predictors are known as **exogenous** variables (X-ogenous to me)
- Outcomes are known as **endogenous** variables (IN-dogenous to me)
- Variables that are both at once are called endogenous variables

<u>Our previous example model:</u>
2 exogenous variables (x1 and x2)
2 endogenous variables (y1 and y2)



<u>Our modified example model:</u>
2 exogenous variables (x1 and x2)
2 endogenous variables (y1 and y2)



New (and Confusing) Terminology

- What parameters get estimated for exogenous "predictor" and endogenous "outcome" variables differs importantly by program!
  - Only the intercepts, residual variances, and residual covariances of "outcome" variables are estimated as part of the likelihood…
- But this distinction is not as clear-cut as one might think…

- By default **in Mplus**, *truly* exogenous predictor variables cannot have missing data (same as in any linear model)
  - Cases with missing predictors are **listwise deleted** out of the model (incomplete data are missing completely at random)
  - Because predictors are not explicitly part of likelihood function
    - LL contains $\hat{y}_i$ for each person and σ_e^2 for each outcome
    - So LL can't be calculated without the predictors that create each $\hat{y}_i$
  - **But these exogenous predictors do not have distributions**…
    - Good when you want to include non-normally-distributed predictors!

"Predictors" as Endogenous Outcomes

- Mplus allows you to bring exogenous predictors into the likelihood → predictors then become "outcomes" in terms of their parameters (estimated means, variances, and covariances)
  - Even if nothing predicts the predictor (it's not really an outcome)
  - These predictors can then have missing data assuming missing at random (conditionally random given the rest of the model)
  - **These predictors then have distributional assumptions (usually MVN)**
  - Mplus will not let endogenous "predictors" have other distributions (so you will have to make them an outcome of something else to fix this)

- **Exogenous predictors are forced into the likelihood in STATA SEM and SAS CALIS** (and I have not been able to find how to force predictors out of the likelihood in those programs)
  - STATA's "xconditional" computes their means, variances, and covariances from the observed data to save time given complete data (and searches for them as model parameters otherwise), but these values then go into the likelihood, which means exogenous predictors have assumed distributions

Model Identification
(assuming all variables are in the likelihood)

- Identification: can the model parameters actually e "solved for"?
  - Requires that # of estimated parameters is ≤ # of possible parameters
  - # possible is sum of # means, variances, and covariances for v variables
    → shortcut formula = possible degrees of freedom = $(v[v + 1] / 2) + v$

- 3 possible model identification scenarios:

  - **Under-identified:** # estimated parameters > # possible → negative df
    - Model is not solvable (parameter estimates cannot be found); game over

  - **Just-identified:** # estimated parameters = # possible → 0 df
    - Model is solvable (is most common scenario perfectly reproduces original data)
    - Assessment of absolute model fit will NOT be relevant (which is a good thing)

  - **Over-identified:** # estimated parameters < # possible → positive df
    - Model is still solvable (and is more parsimonious description of original data)
    - Assessment of absolute model fit is then necessary (more relevant for latent variables)

What Goes In (data used as input)

- Observed mean per variable
- Observed variance per variable
- Observed covariance between each pair of variables
- This is the data the model is trying to "fit"!

What Comes Out (estimated parameters)

- Estimated intercept per variable (to *perfectly* re-create the observed variable means)
- Estimated residual variance per variable (to *perfectly* re-create the observed variances)
- Estimated regression path or covariance between each pair of variables (to predict their observed covariances)
  - If some are omitted, then observed covariances will not be perfectly reproduced → **room for misfit**

Estimated Parameters and Model Fit

- If fewer than all possible parameters are estimated, then assessment of **absolute model fit** is needed: how well do the model-predicted parameters match the corresponding estimates from the original data?
  - I would recommend sticking with "just-identified" path models (# estimated parms= # possible parms) so that fit is not an issue

- Absolute model fit is assessed with a specific variant of the likelihood ratio test for relative fit you already know…
  - In fact, we did this for repeated measures data: when testing whether an unstructured R matrix (of all possible separately estimated variances and covariances) fit better than a simpler alternative model for the variance…
    - Did a single variance and covariance (compound symmetry) adequately predict all possible UN separately estimated variances and covariances?

Review of Likelihood Ratio Tests

- Multivariate models require assessment of **relative model fit**: how well does the model fit relative to other possible models?

- Relative fit is indexed by overall model **log-likelihood** (**LL**):
  - Log of likelihood for each person's outcomes given model parameters
  - Sum log-likelihoods across all independent persons = **model LL**

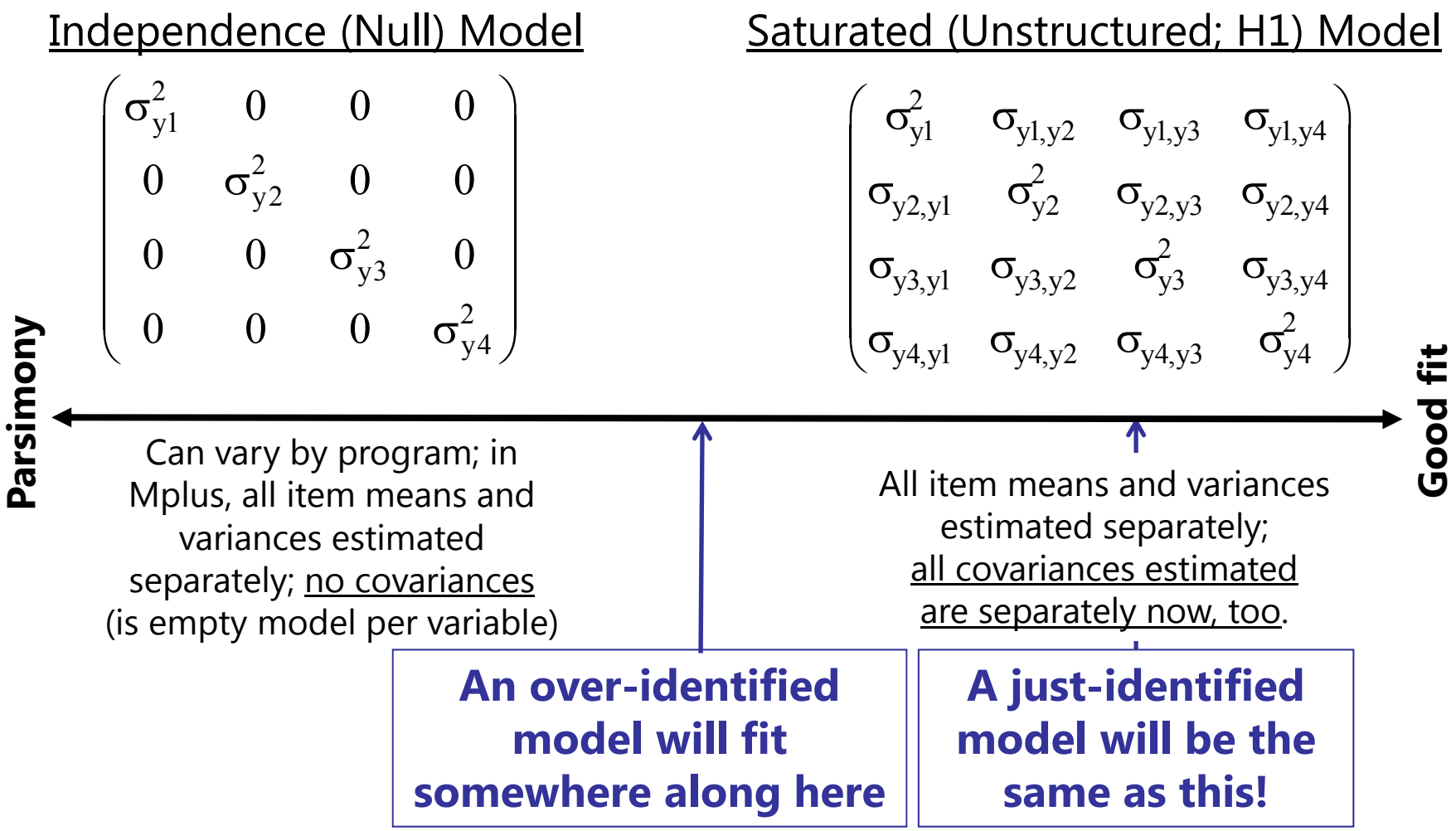
- Nested models are compared using a **"likelihood ratio test": −2ΔLL test** (aka, "χ^2 test" in SEM and path models)

"fewer" = from model with fewer parameters
"more" = from model with more parameters

Results of 1. & 2. must be positive values!

1. Calculate **−2ΔLL:** if given −2LL, do $-2\Delta LL = (-2LL_{fewer}) - (-2LL_{more})$
   if given LL, do $-2\Delta LL = -2 * (LL_{fewer} - LL_{more})$
2. Calculate **ΔdF =** $(\#\ Parms_{more}) - (\#\ Parms_{fewer})$
3. **Compare −2ΔLL to χ^2 distribution with df = Δdf**
4. Get *p*-value from CHIDIST in excel or LRTEST option in STATA

Baselines for Assessing Model Fit
(Item means are saturated via intercepts in both)



LRT for comparison with saturated model is already given in your output

Mplus output:

```
MODEL FIT INFORMATION
Number of Free Parameters                       13

Loglikelihood
          H0 Value                       -3633.619
          H1 Value                       -3626.022
Chi-Square Test of Model Fit
          Value                             15.192
          Degrees of Freedom                     1
          P-Value                           0.0001
```

SAS PROC CALIS output (as −2LL):

Saturated Model -2 Log Likelihood		7252.0448
Absolute Index	Fit Function	11.1290
	-2 Log-Likelihood	7267.2371
	Chi-Square	15.1924
	Chi-Square DF	1
	Pr > Chi-Square	<.0001

H1 Saturated (Unstructured) Model

$$\begin{pmatrix} \sigma^2_{y1} & \sigma_{y1,y2} & \sigma_{y1,y3} & \sigma_{y1,y4} \\ \sigma_{y2,y1} & \sigma^2_{y2} & \sigma_{y2,y3} & \sigma_{y2,y4} \\ \sigma_{y3,y1} & \sigma_{y3,y2} & \sigma^2_{y3} & \sigma_{y3,y4} \\ \sigma_{y4,y1} & \sigma_{y4,y2} & \sigma_{y4,y3} & \sigma^2_{y4} \end{pmatrix}$$

"Model fit" χ^2 is from a $-2\Delta LL$ test of your **H0** model vs. saturated **H1** model

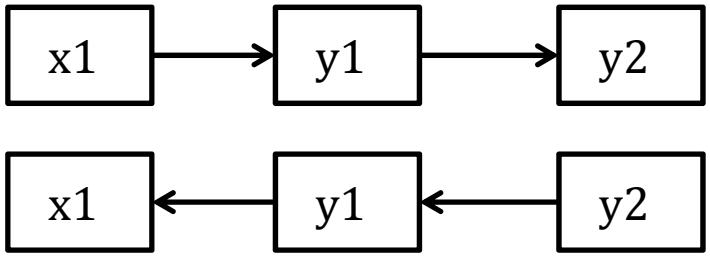
STATA SEM output:

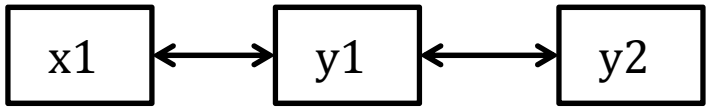
```
Structural equation model
Estimation method  = mlmv
Log likelihood     = -3633.6186

Likelihood ratio
 chi2_ms(1) |  15.192   model vs. saturated
   p > chi2 |   0.000
```

Model Identification Examples

(in which each variable has an estimated mean/intercept and variance/residual variance)

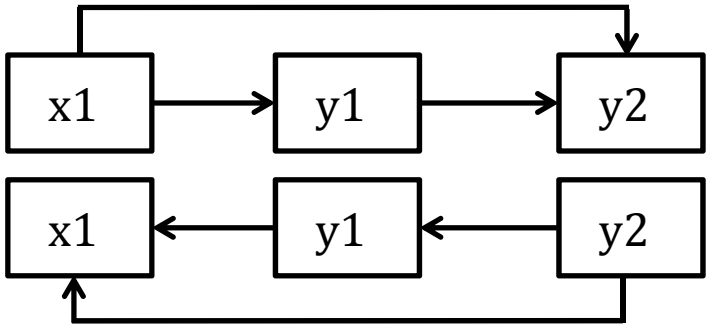
- Over-identified: have positive df leftover (estimated<possible)

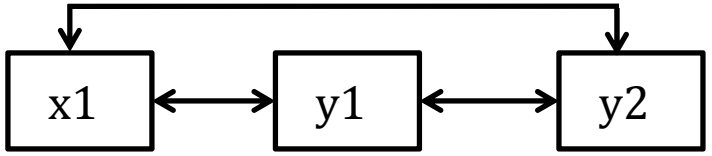




These 3 models all have equivalent fit with **df=1** (for the 1 missing direct relationship).

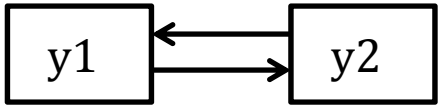
- Just-identified: have 0 df leftover (estimated=possible)





These 3 models all have equivalent fit with df=0 (for 0 missing direct relationships).

- Under-identified: have negative df (estimated>possible)



This model is trying to estimate 2 paths using only 1 covariance (can't be solved).

3 Steps in Assessing Model Fit for Over-Identified Models

1. Global absolute model fit
   - *Does the model 'work' as a whole?*

2. Local absolute model fit
   - *Are there any more specific problems?*

3. Revise the model as needed until absolute fit is achieved
   - *Add parameters? Absolute model fit can be better or not better*
   - *Remove parameters? Absolute model fit can be worse or not worse*

---

4. For all models: Interpret parameters and consider effect size
   - *Do the numbers make sense? Are they useful?*
   - *A good-fitting model IS not the same as a GOOD MODEL!*

Indices of Global Absolute Model Fit

- Primary fit index: obtained model $\boldsymbol{\chi^2 = 2 * N * FML}$
  - χ^2 is evaluated based on model df (# parameters left over)
  - Tests null hypothesis that $\boldsymbol{\Sigma = S}$ (that model is perfect), so **significance is bad** (i.e., smaller χ^2, bigger *p*-value is better)
  - Just using χ^2 to index model fit is usually insufficient, however:
    - Obtained χ^2 depends largely on sample size (N)
    - Is unreasonable null hypothesis (perfect fit, really??)
    - Only possible given balanced data (as typical in SEM and path models)

- 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

Indices of Global Absolute Model Fit

- Absolute Fit: $\boldsymbol{\chi^2}$
  - Don't use "ratio rules" like $\chi^2/\mathrm{df} > 2$ or $\chi^2/\mathrm{df} > 3$

- Absolute Fit: **SRMR**
  - **Standardized Root Mean Square Residual**
  - Get difference of standardized $\boldsymbol{\Sigma}$ and $\mathbf{S}$ → residual matrix
  - Sum the squared residuals of the correlation matrix across items, divide by number of residuals (i.e., matrix elements)
  - Ranges from 0 to 1: smaller is better
  - ".08 or less" → good fit

- See also: **RMR** (**Root Mean Square Residual**)

Indices of Global Absolute Model Fit

Parsimony-Corrected: **RMSEA**

- **Root Mean Square Error of Approximation**
- Relies on a "non-centrality parameter" (NCP)
  - Indexes how far off your model is → χ^2 distribution shoved over
  - NCP → $\mathrm{d} = (\chi^2 - \mathrm{df}) / \mathrm{N}$ Then, $\mathrm{RMSEA} = \sqrt{\mathrm{d}/\mathrm{df}}$

- 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

Indices of Global Absolute Model Fit

Comparative (Incremental) Fit Indices

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

- **CFI: Comparative Fit Index**
  - Also based on idea of NCP ($\chi^2 - df$)
  - $CFI = 1 - \frac{\max[(\chi^2_T - df_T), 0]}{\max[(\chi^2_T - df_T), (\chi^2_N - df_N), 0]}$ T = target model, N = null model
  - From 0 to 1: bigger is better, > .90 = "acceptable", > .95 = "good"

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

4 Steps in Model Evaluation

2. Identify localized model strain

- Global model fit means that the observed and predicted item covariance matrices aren't too far off on the whole… this says nothing about the specific covariances to be predicted
- Should inspect **normalized model residuals** for that → Local fit
  - Available via **RESIDUAL** output option in Mplus, RESIDUAL=NORM on PROC CALIS statement in SAS, or "estat gof, stats(all)" in STATA
  - Normalized as residual/SE → **works like a z-score**
  - Relatively large absolute values indicate "localized strain"
  - **Positive** residual → Variables are more related than you predicted
  - **Negative** residual → Variables are less related than you predicted

- Should add relationships to fix local model fit
  - Can test new paths and covariances via univariate Wald tests or likelihood ratio test (allowed given use of regular-flavor ML in path models)

Summary: Steps 1, 2, and 3

1. **Assess global absolute model fit**
   - Recall that variable means and variances are perfectly predicted (just-identified) → *misfit comes from messed-up covariances*
   - χ^2 is sensitive to large sample size, so pick at least one global fit index from each class; hope they agree (e.g., CFI, RMSEA)
2. **Identify localized model strain**
   - Global model fit means that the observed and predicted covariance matrices aren't too far off on the whole… says nothing about the specific matrix elements (reproduction of each covariance)
   - Consider normalized residuals and modification indices to try and "fix" the model – add missing relationships that should be there
3. **Revise the model until it fits**

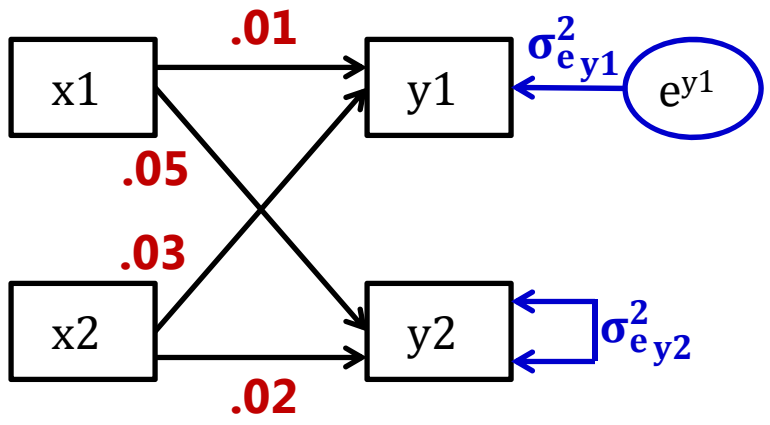
Good global and local fit? Great, but we're not done yet…

4 Steps in Model Evaluation

4. Inspect **parameter effect sizes** and significance
   - A good-fitting model does not necessarily imply a good model!
     - Can reproduce lack of covariance quite well and still not have anything useful – e.g., correlation of .2 → 4% shared variance?
     - **Effect size (R^2 for variance explained) is practical significance**

This example model may have "excellent fit" (testable because df=2) but no significant regression paths…

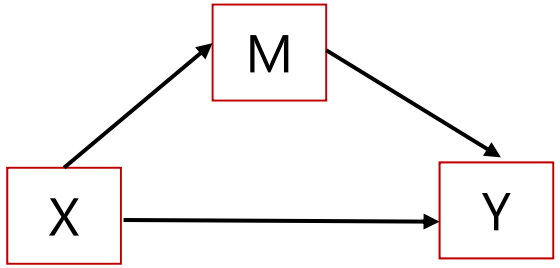
Why? Good absolute fit just means it has successfully reproduced the (non)relationships among these variables—not whether there are relationships worth reproducing!



Terminology: Mediation ≠ Moderation

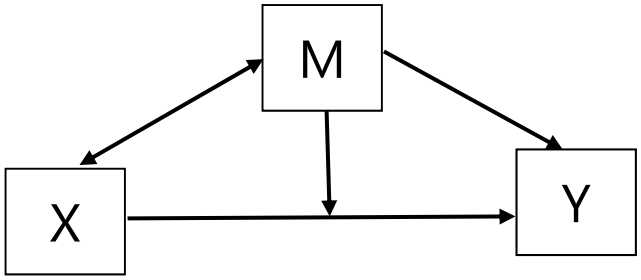
Mediational model (regression with better marketing):

- X **causes** M, M **causes** Y
- M is an outcome of X but a predictor of Y

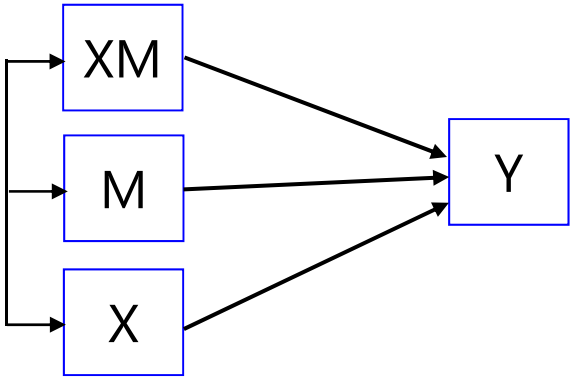


Moderator model:

- M adjusts the size of X→Y relationship
- M is a predictor of Y, and is **correlated** with X
- Moderation is represented by an **interaction** effect



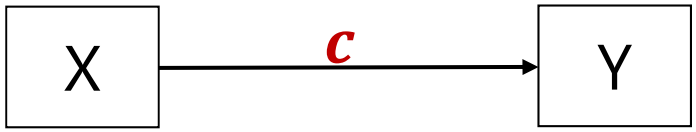
This figure does NOT depict an estimable model.

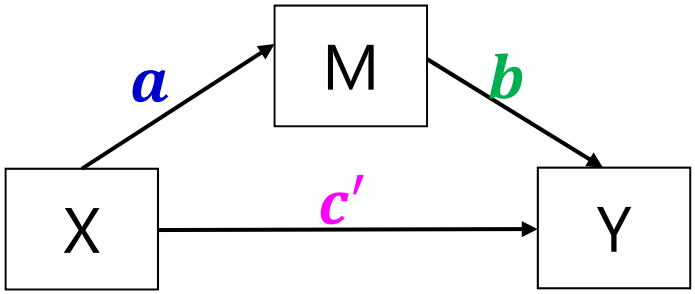


This is what is actually implied by above model.

Terminology: Mediation Effects

c = uncontrolled X to Y path
(Y regressed on X)





The big question in mediation:

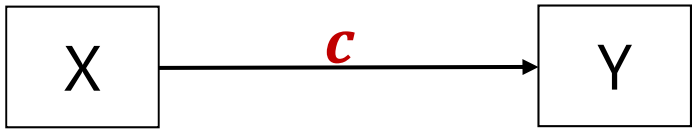
- Phrased as usual regression → *Is the effect of X predicting Y still significant after controlling for M?*
- Phrased as "mediation" → *Is the effect of X predicting Y significantly mediated by M?* **OR** *Is there a significant indirect effect of X through M in predicting Y?*
- Phrased either way, is $c \neq c'$?

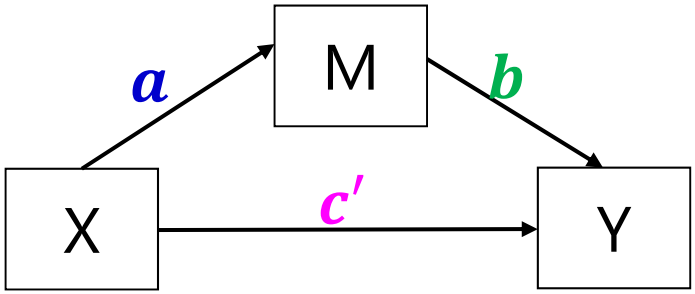
Direct Effects:

- a = X to M path (M on X;)
- b = M to Y path (Y on M;)
- c' = X to Y path controlled for M (Y on X;)
- $a * b$ = indirect effect of X to Y
- The estimates for $c - c'$ and $a * b$ will be equivalent in MVN observed variables (if same *N*)

Old versus New Rules for Mediation

c = uncontrolled X to Y path
(Y regressed on X)





- Baron & Kenny (1986, JPSP) rules were standard for a long time…
  - Simulation studies have found these rules to be way too conservative
- Old rule that can now be broken:
  - X must predict Y in the first place (c must be initially significant)
  - When not? Differential power for paths; suppressor effects of mediators
  - Mediation is really about whether $c \neq c'$, not whether each is significant
- Old rules that pry still hold:
  - X must predict M (a must be significant)
  - M must predict Y (b must be significant)

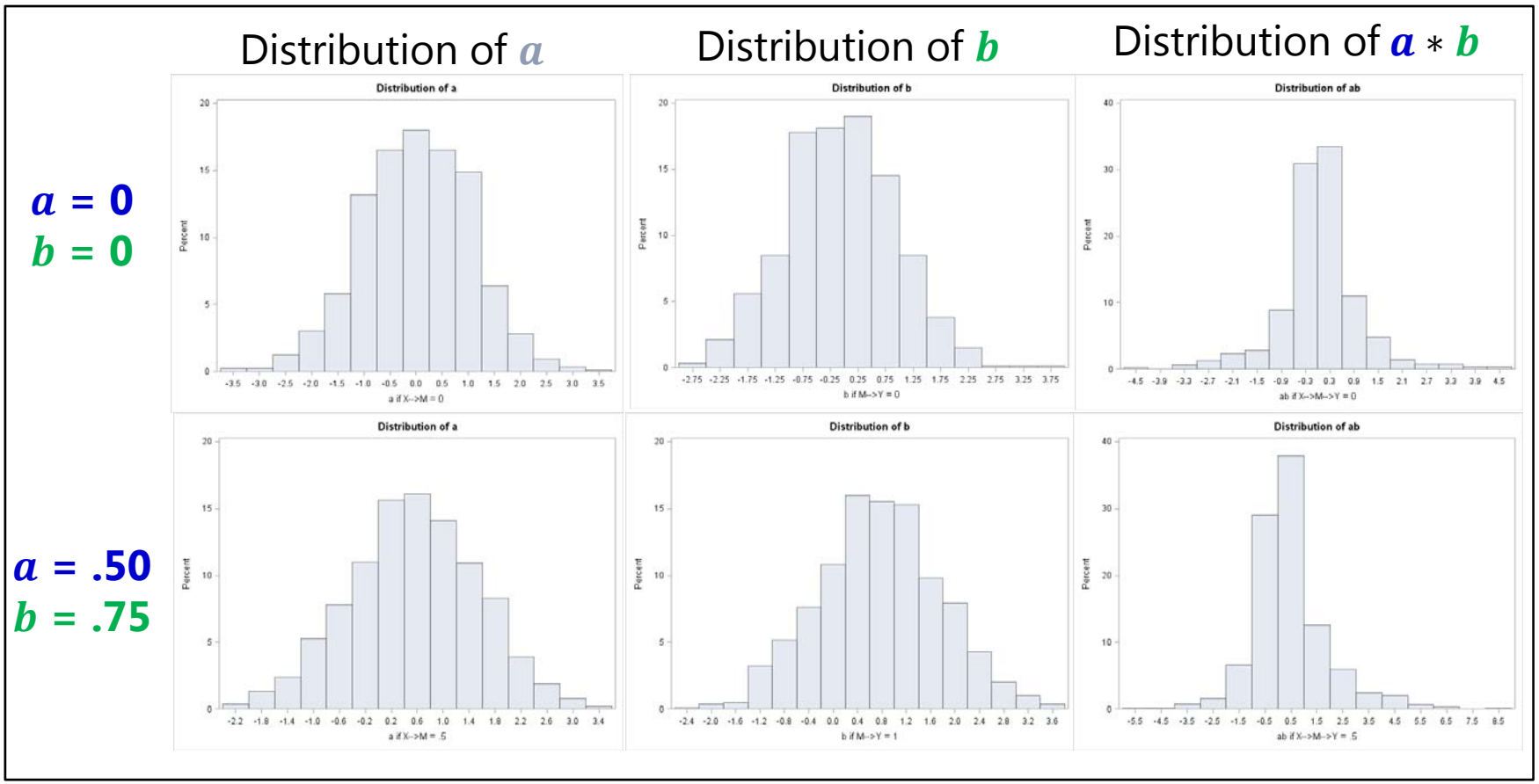
Testing Significance of Mediation

- Need to obtain a SE in order to test if $\boldsymbol{c - c' = 0}$ or if $\boldsymbol{a * b = 0}$
  - For $\boldsymbol{c - c'}$ → "difference in coefficients SE"
  - For $\boldsymbol{a * b}$ → "product of coefficients SE" → we'll start here

- Use "multivariate delta method" (second-derivative approximation shown here) to get SE for product of two random variables $\boldsymbol{a * b}$
  - $SE_{a*b} = \sqrt{a^2 SE_b^2 + b^2 SE_a^2 + SE_a^2 SE_b^2}$
  - An equivalent formula to calculate SE_{a*b} that may have less rounding error because it avoids squaring $\boldsymbol{a}$ and $\boldsymbol{b}$ is $SE_{a*b} = \frac{ab\sqrt{t_a^2 + t_b^2 + 1}}{t_a t_b}$
  - This is known as the "Sobel test" and can be calculated by hand using the results of a simultaneous path model or separate regression models, also provided through MODEL INDIRECT/CONSTRAINT in Mplus, NLCOM in STATA SEM, or TESTFUNC in SAS PROC CALIS

Testing Significance of Mediation

- One problem: we *shouldn't* use this SE for usual significance test
  - So, nope: $t_{indirect} = \frac{a*b}{SE_{a*b}}$ or $95\%\ CI = a * b \pm 1.96 * SE_{a*b}$
  - Why? Although the estimates for a and b will be normally distributed, the estimate of their product won't be, especially if a and b are near 0



Testing Significance of Mediation

- So what do we do? Another idea based on same premise:
  - For $\boldsymbol{a} * \boldsymbol{b}$ → find "distribution of the product SE" → $\boldsymbol{z_a} * \boldsymbol{z_b} = \frac{\boldsymbol{a}}{\boldsymbol{SE_a}} * \frac{\boldsymbol{b}}{\boldsymbol{SE_b}}$ in which the sampling distribution does not have a tractable form, but tables of critical values have been derived through simulation for the single mediator case (but may not generalize to complex models)
  - Implemented in PRODCLIN program for use with SAS, SPSS, and R

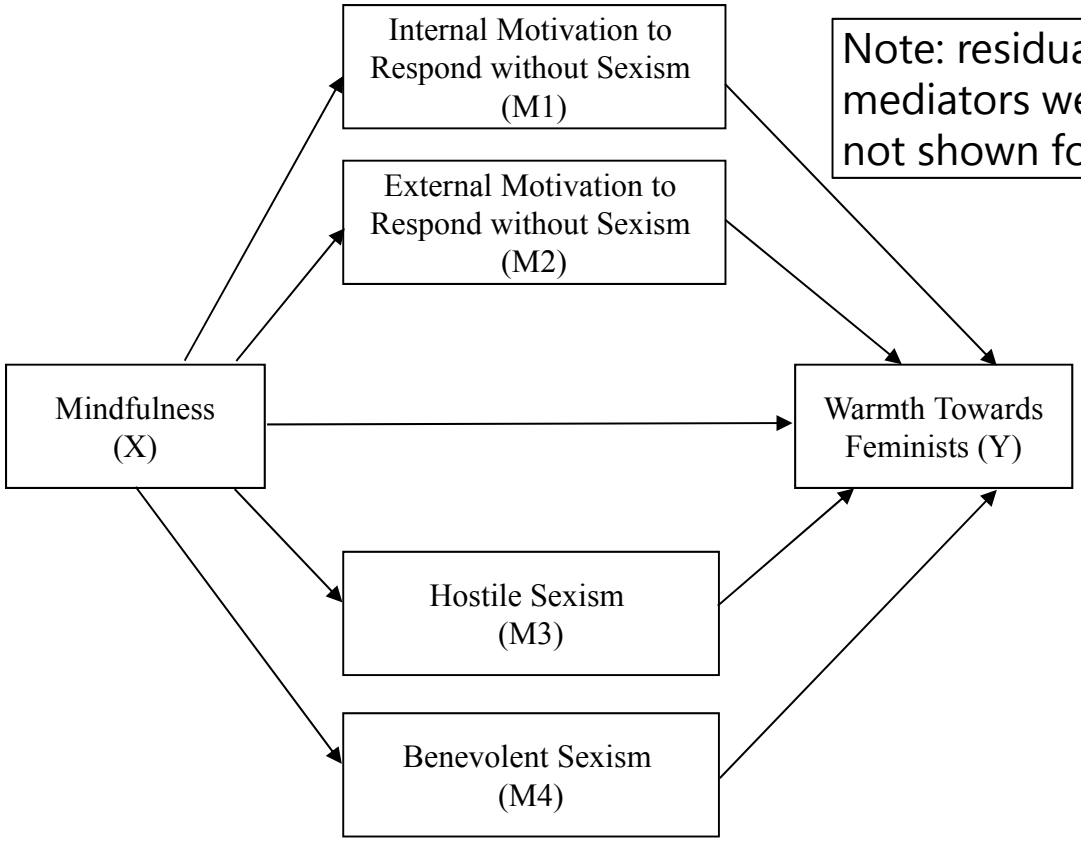
- A better solution: **bootstrap the data** to find the empirical SE and asymmetric CI for the indirect effect
  - Bootstrap = draw n samples with replacement from your **data**, re-estimate mediation model and calculate $\boldsymbol{a} * \boldsymbol{b}$ within each bootstrap sample
  - Point estimate of $\boldsymbol{a} * \boldsymbol{b}$ is mean or median over n bootstrap samples
  - SE_{a*b} is standard deviation of estimated $\boldsymbol{a} * \boldsymbol{b}$ over n bootstrap samples
  - 95% CI can be computed as estimates at the 2.5 and 97.5 percentiles
  - Typically at least 500 or 1000 n bootstrap samples are used

Testing Significance of Mediation

- There are multiple kinds of bootstrap CIs possible in testing the significance of the $a * b$ indirect effect within MVN data
  - Regular bootstrap CI = "**percentile**" (as just described)
    - In Mplus, OUTPUT: CINTERVAL(bootstrap); in STATA SEM, vce(bootstrap)
  - **Bias-corrected bootstrap** CI = shifts CIs so median is sample estimate *** *Supposed to be best one*
    - In Mplus, OUTPUT: CINTERVAL(BCbootstrap); not sure about STATA SEM
  - Accelerated bootstrap CI = ???
    - Not given in Mplus (as far as I know); not sure about STATA SEM

- For not simply MVN data (i.e., non-normal mediators or outcomes, multilevel data), a different bootstrap approach can be used as a separate step using any program's output
  - *Parametric*, *Monte Carlo*, or *empirical-M* bootstrap → Draw repeatedly from a and b parameter distributions instead of the data, then compute point estimates, SE, and CIs from those distributions
  - See http://www.quantpsy.org/medn.htm for online calculators

Our Mediation Example



Note: residual covariances among mediators were estimated but are not shown for diagram clarity.

Figure 1 from: Gervais, S. J. & Hoffman, L. (2013). Just think about it: Mindfulness, sexism, and prejudice towards feminists. *Sex Roles, 68*(5), 283-295.

Research questions:

(1) To what extent do these four mediators account for the relationship between mindfulness and warmth towards feminists?

(2) How do these direct and indirect effects differ by gender?

Path Models and Mediation: Summary

- Path models are a very useful way to examine many different multivariate hypotheses simultaneously:
  - Unique direct and indirect effects ("mediation")
  - Differences in effect size (via model constraints)
  - Relationships among mediators or outcomes (direct and indirect effects)

- Good fit is a pre-requisite to actually interpreting the model results, but good fit does *not* mean it is a good model
  - Good fit = model reproduces the covariance matrix of the likelihood variables (but it does not indicate how big or small those relationships are)
  - However – when all possible relationships among variables are estimated (either as covariances or direct regressions), fit is perfect and irrelevant
    - We used to call this "multivariate regression" with an "unstructured R matrix"

- Watch out for assumptions about exogenous predictor variables
  - Are their means, variances, and covariances part of the likelihood? Then they have an assumed distribution (usually MVN), which may not make any sense!