Latent Variable Models with Blimp

Brian T. Keller

Assistant Professor
Statistics, Measurement, & Evaluation in Education
Educational, School, & Counseling Psychology
University of Missouri

Developer, Blimp

Educational, School and
Counseling Psychology
University of Missouri

Blimp Software

- General-purpose Bayesian estimation for regression and path models.
- Single and multilevel models
- Allows for latent variables, incomplete predictors and outcomes
- Interactive and nonlinear effects
- Nonnormal data
- And more!

Available at

`https://www.appliedmissingdata.com/blimp`

Workshop Content

Blimp available at

`https://www.appliedmissingdata.com/blimp`

Workshop content at

`https://github.com/blimp-stats/Yonsei-Workshop`

Acknowledgements

Craig Enders, Ph.D. (UCLA)

Han Du, Ph.D. (UCLA)

Jiwon Kim, Ph.D. (UT Austin)

Suyoung Kim, M.A. (UT Austin)

IES Institute of Education Sciences

YONSEI UNIVERSITY 연세대학교 1885

Workshop Outline

Day 1:

- An Introduction to Missing Data and Bayesian Statistics with Blimp

Day 2:

- Finish remaining material from Day 1.
- Latent Variable Models with Blimp

Day 1's Review

- Missing Data Processes
- Introduction to Bayesian Statistics
- Fitting Regression Models in Blimp
- Understanding Blimp Output
- Incomplete Categorical Variables
- Interaction Effects in Blimp

Day 2's Overview

- Latent Variable Modeling with Blimp
- Latent Interactions with Blimp
- Overview of Estimation and Latent Scores
- Evaluating Latent Interactions
- Multilevel Analysis Example
- Parallel Growth Process Example

$$Y = \beta_0 + \beta_1 M + \beta_2 X + \beta_3 (M \times X) + \epsilon$$

$$\eta_Y = \beta_0 + \beta_1\eta_M + \beta_2\eta_X + \beta_3(\eta_M \times \eta_X) + \zeta$$

Latent Variable Modeling with Blimp

Factored Regression Models

Suppose, we are interested in modeling a three variable problem with X, Y, and Z.

Their joint distribution are represented symbolically as:

$$f(X, Y, Z)$$

Often, we may assume that all three variables follow a multivariate normal distribution with some unknown mean vector ($\boldsymbol{\mu}$) and covariance matrix ($\boldsymbol{\Sigma}$):

$$f(X, Y, Z) = \mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

As an alternative to modeling the joint distribution, we can model this density using conditional models.

By applying the Chain Rule of Probability, we ***factorize*** the distribution into three univariate distributions:

$$f(X, Y, Z) = f\left(X \mid Y, Z\right) \times f\left(Y \mid Z\right) \times f(Z)$$

We then specify models for these three distributions. For example, a saturated covariance matrix can be modeled via three linear regressions:

$$f\left(X \mid Y, Z\right) \rightarrow \quad X \sim \mathcal{N}\left(\alpha_0 + \alpha_1 Y + \alpha_2 Z, \sigma^2_{rX}\right)$$

$$f\left(Y \mid Z\right) \rightarrow \quad Y \sim \mathcal{N}\left(\beta_0 + \beta_1 Z, \sigma^2_{rY}\right)$$

$$f(Z) \rightarrow \quad Z \sim \mathcal{N}\left(\gamma_0, \sigma^2_{Z}\right)$$

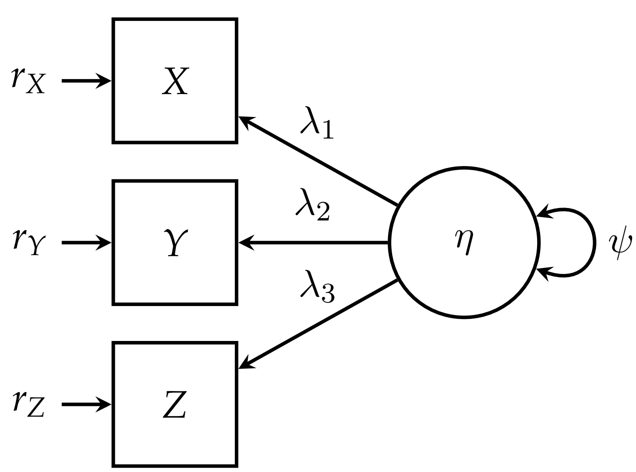
Note: We are using the regression coefficients as a tool to maintain the associations among the variables.

For example, the β_1 coefficient is a ***reparameterization*** of the covariance (σ_{YZ}):

$$\sigma_{YZ} = \beta_1 \sigma_Z^2$$

Measurement Models as Factored Regressions

Suppose our trivariate example loads onto a single latent factor:



The common factor model assumes a multivariate normal distribution, with a model implied mean structure and covariance matrix:

$$f(X, Y, Z) = \mathcal{N}_3\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)$$

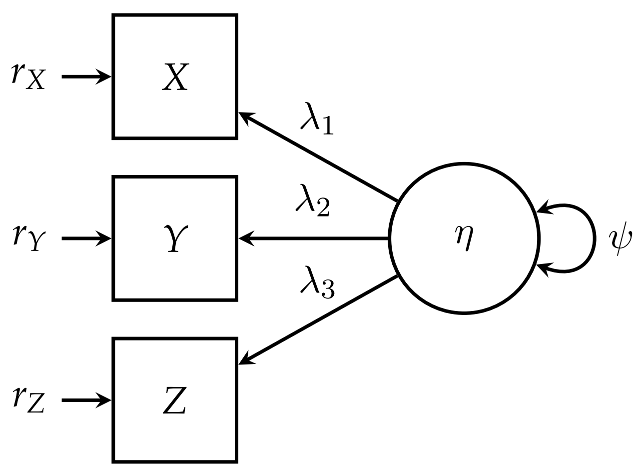
$$\widehat{\boldsymbol{\mu}} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha} \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^\intercal + \boldsymbol{\Theta}$$

The factored regression approach explicitly focuses on the joint distribution of the three indicators and latent factor:

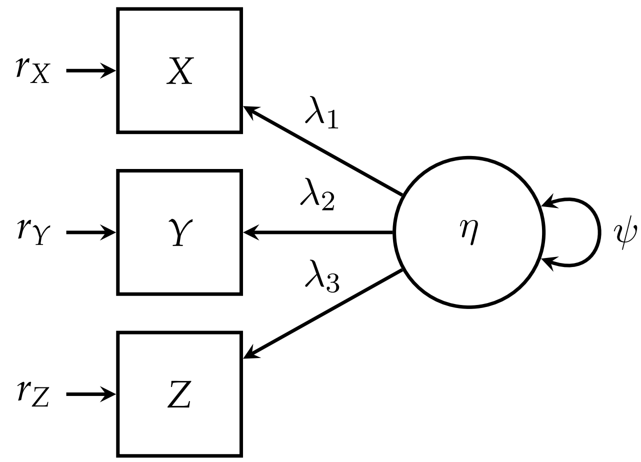
$$f(X, Y, Z, \eta)$$

We fully factored the multivariate distribution into a product of four univariate distributions:

$$f(X, Y, Z, \eta) = f\big(X \mid Y, Z, \eta\big) \times f\big(Y \mid Z, \eta\big) \times f\big(Z \mid \eta\big) \times f(\eta)$$



$$f\big(X \mid Y, Z, \eta\big) \times f\big(Y \mid Z, \eta\big) \times f\big(Z \mid \eta\big) \times f(\eta)$$



$$f\left(X \mid \eta\right) \times f\left(Y \mid \eta\right) \times f\left(Z \mid \eta\right) \times f(\eta)$$

$$f\left(X \mid \eta\right) \rightarrow \qquad x_i = \nu_1 + \lambda_1\eta_i + r_{Xi}$$

$$f\left(Y \mid \eta\right) \rightarrow \qquad y_i = \nu_2 + \lambda_2\eta_i + r_{Yi}$$

$$f\left(Z \mid \eta\right) \rightarrow \qquad z_i = \nu_3 + \lambda_3\eta_i + r_{Zi}$$

$$f(\eta) \rightarrow \qquad \eta_i = \alpha + \zeta_i$$

What if my indicators are categorical?

$$f\big(X \mid \eta\big) \rightarrow \qquad \text{probit}(x_i) = \nu_1 + \lambda_1\eta_i + r_{Xi}$$

$$f\big(Y \mid \eta\big) \rightarrow \qquad \text{probit}(y_i) = \nu_2 + \lambda_2\eta_i + r_{Yi}$$

$$f\big(Z \mid \eta\big) \rightarrow \qquad \text{probit}(z_i) = \nu_3 + \lambda_3\eta_i + r_{Zi}$$

$$f(\eta) \rightarrow \qquad \eta_i = \alpha + \zeta_i$$

$$f\big(X \mid \eta\big) \rightarrow \qquad \text{logit}(x_i) = \nu_1 + \lambda_1\eta_i + r_{Xi}$$

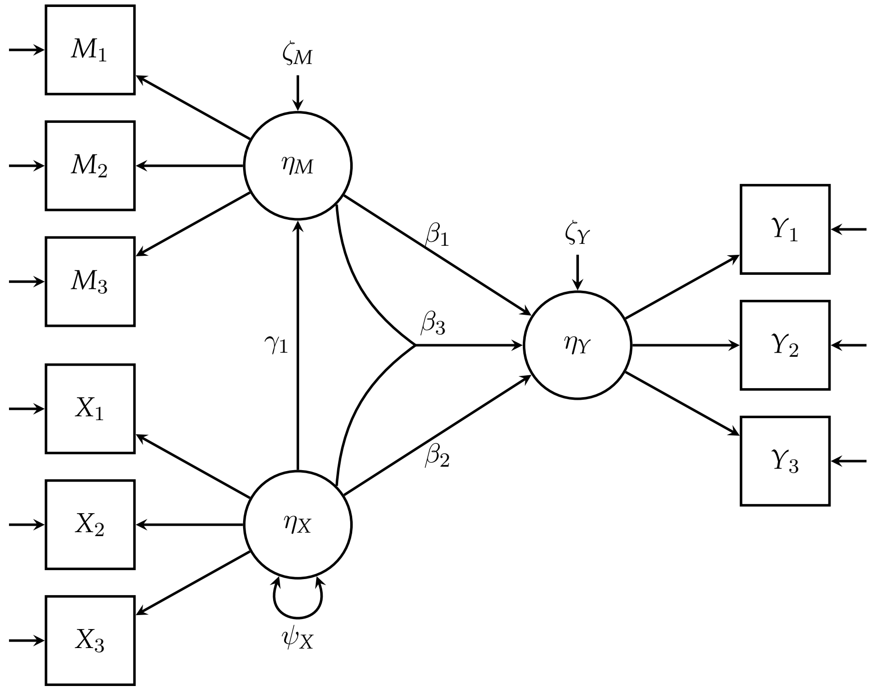
$$f\big(Y \mid \eta\big) \rightarrow \qquad \text{logit}(y_i) = \nu_2 + \lambda_2\eta_i + r_{Yi}$$

$$f\big(Z \mid \eta\big) \rightarrow \qquad \text{logit}(z_i) = \nu_3 + \lambda_3\eta_i + r_{Zi}$$

$$f\big(\eta\big) \rightarrow \qquad \eta_i = \alpha + \zeta_i$$

Latent Interactions in Blimp

Latent Interactions with Blimp

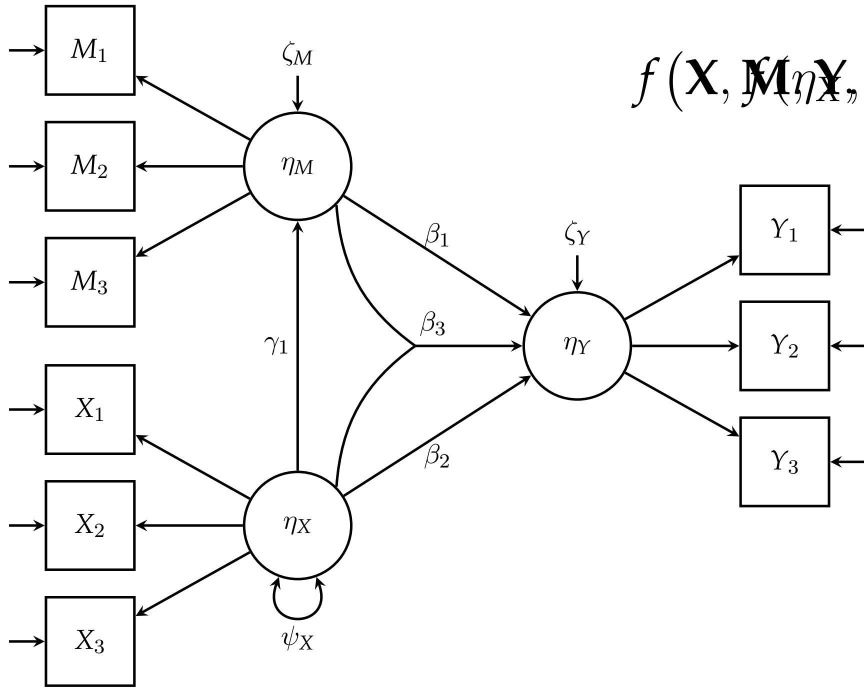


Factor the joint density into measurement and structural densities:

$$f(\mathbf{X}, \mathbf{M}, \mathbf{Y}, \eta_X, \eta_M, \eta_Y) = f\left(\mathbf{X}, \mathbf{M}, \mathbf{Y}, \mid \eta_X, \eta_M, \eta_Y\right) \times f(\eta_X, \eta_M, \eta_Y)$$

Boldface variables represent the set of all indicators:

$$\mathbf{X} = \{X_1, X_2, X_3\} \qquad \mathbf{M} = \{M_1, M_2, M_3\} \qquad \mathbf{Y} = \{Y_1, Y_2, Y_3\}$$



$$f\left(\mathbf{X}, \mathbf{M}, \mathbf{Y} \mid \eta_X, \eta_M, \eta_Y\right)$$

We factor the measurement density into three separate densities:

$$f\left(\mathbf{X}, \mathbf{M}, \mathbf{Y} \mid \eta_X, \eta_M, \eta_Y\right) = f\left(\mathbf{X} \mid \eta_X\right) \times f\left(\mathbf{M} \mid \eta_M\right) \times f\left(\mathbf{Y} \mid \eta_Y\right)$$

Each of the three conditional densities map onto the single factor model.

For example: $f\left(\mathbf{X} \mid \eta_X\right) = f\left(X_1 \mid \eta_X\right) \times f\left(X_2 \mid \eta_X\right) \times f\left(X_3 \mid \eta_X\right)$

Factored Structural Density

We factor the structural density into three separate densities:

$$f(\eta_X, \eta_M, \eta_Y) = f\left(\eta_Y \mid \eta_M, \eta_X\right) \times f\left(\eta_M \mid \eta_X\right) \times f(\eta_X)$$

The form of the models map onto a simple linear regression with an $\eta_M \times \eta_X$ interaction:

$$f\left(\eta_Y \mid \eta_X, \eta_M\right) \rightarrow \quad \eta_{Yi} = \alpha_Y + \beta_1 \eta_{Mi} + \beta_2 \eta_{Xi} + \beta_3 \eta_{Mi} \times \eta_{Xi} + \zeta_{Yi}$$

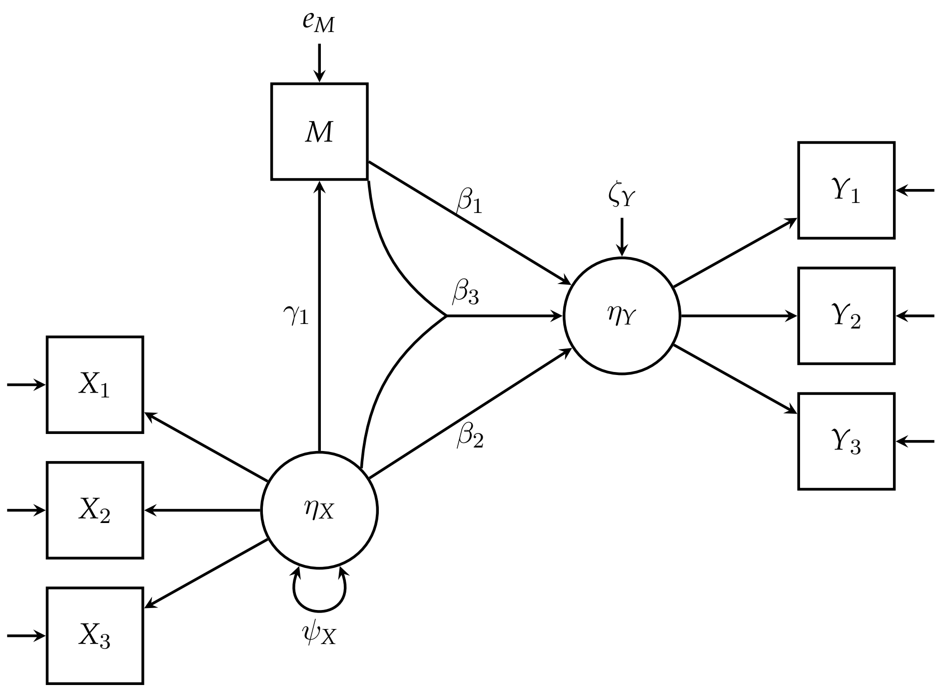
$$f\left(\eta_M \mid \eta_X\right) \rightarrow \quad \eta_{Mi} = \alpha_M + \gamma_1 \eta_{Xi} + \zeta_{Mi}$$

$$f(\eta_X) \rightarrow \quad \eta_{Xi} = \alpha_X + \zeta_{Xi}$$

Blimp Syntax

```
LATENT:
    eta_x eta_m eta_y;

MODEL:
    structural_model: # Label for organization
        eta_y ~ eta_x eta_m eta_x*eta_m;
        eta_m ~ eta_x;
        eta_x ~ intercept@0;

    measurement_model: # Label for organization
        eta_x -> x1@1 x2 x3;
        eta_m -> m1@1 m2 m3;
        eta_y -> y1@1 y2 y3;
```


What if M is an observed variable?



The factorization of the structural model barely changes:

$$f\left(\eta_Y \mid \eta_X, M\right) \rightarrow \quad \eta_{Yi} = \alpha_Y + \beta_1 m_i + \beta_2 \eta_{Xi} + \beta_3 (m_i \times \eta_{Xi}) + \zeta_{Yi}$$

$$f\left(M \mid \eta_X\right) \rightarrow \quad m_i = \alpha_M + \gamma_1 \eta_{Xi} + r_{Mi}$$

$$f(\eta_X) \rightarrow \quad \eta_{Xi} = \alpha_X + \zeta_{Xi}$$

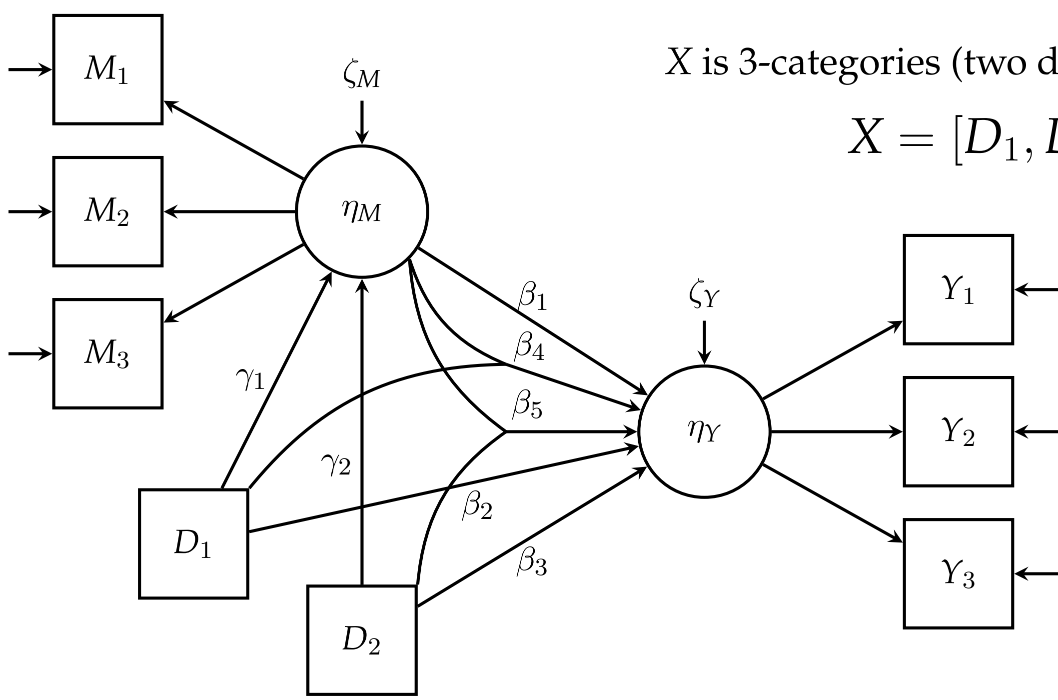
What if X is a nominal variable?

For example, X is 3-categories (two dummy codes):

$$X = [D_1, D_2]$$

X is 3-categories (two dummy codes):

$$X = [D_1, D_2]$$



If X is fully observed:

$$f\left(\eta_Y \mid X, M\right) \rightarrow \qquad \eta_{Yi} = \alpha_Y + \beta_1 m_i + \beta_2 d_{1i} + \beta_3 d_{2i} + \beta_4(m_i \times d_{1i}) + \beta_5(m_i \times d_{2i}) + \zeta_{Yi}$$

$$f\left(M \mid X\right) \rightarrow \qquad m_i = \alpha_M + \gamma_1 d_{1i} + \gamma_2 d_{2i} + r_{Mi}$$

If X is ***incomplete***:

$$f\left(\eta_Y \mid X, M\right) \rightarrow \qquad \eta_{Yi} = \alpha_Y + \beta_1 m_i + \beta_2 d_{1i} + \beta_3 d_{2i} + \beta_4(m_i \times d_{1i}) + \beta_5(m_i \times d_{2i}) + \zeta_{Yi}$$

$$f\left(M \mid X\right) \rightarrow \qquad m_i = \alpha_M + \gamma_1 d_{1i} + \gamma_2 d_{2i} + r_{Mi}$$

$$f(X) \rightarrow \qquad \text{probit}\left(\begin{bmatrix} d_{1i} \\ d_{2i} \end{bmatrix}\right) = \begin{bmatrix} \alpha_{D1} \\ \alpha_{D2} \end{bmatrix} + \begin{bmatrix} e_{D1i} \\ e_{D2i} \end{bmatrix}$$

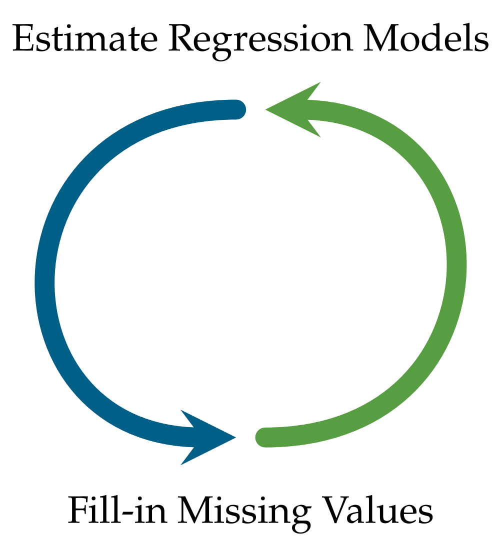
Overview of Estimation and Latent Scores

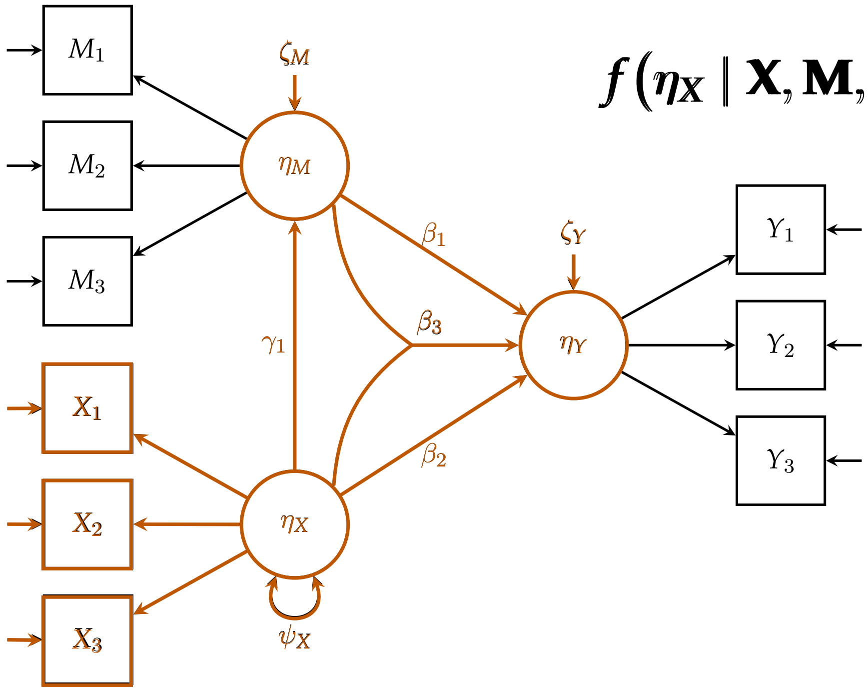
Markov chain Monte Carlo (MCMC Estimation

Do for $t = 1$ to T iterations:

- Sampling the parameters conditional on the observed and imputed data
- Sampling the missing observations conditional on observed data and previously sampled parameters

Repeat





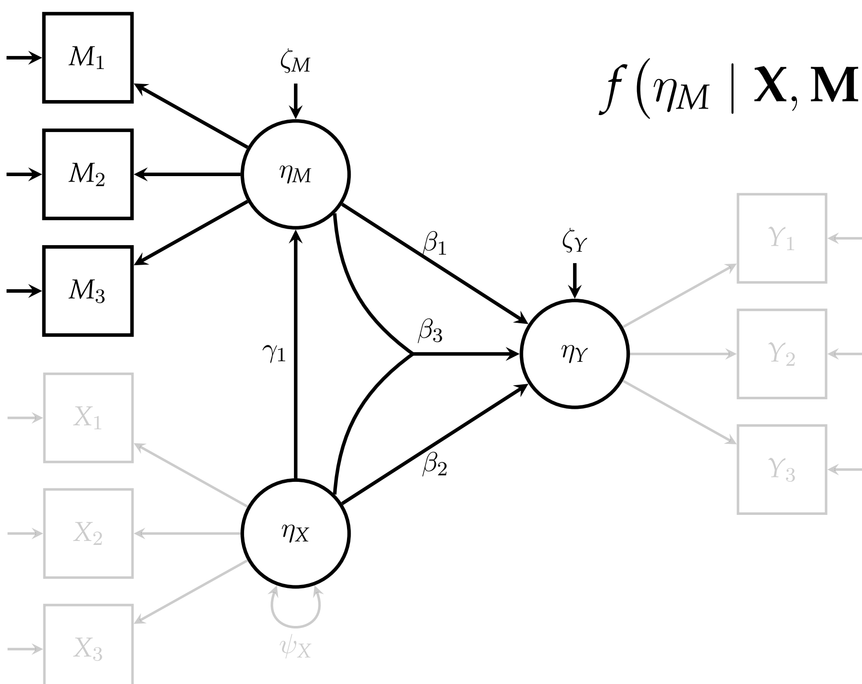
$$\boldsymbol{f\left(\eta_X \mid \mathbf{X}, \mathbf{M}, \mathbf{Y}, \eta_Y, \eta_M\right)} \propto$$

$$f\left(\eta_X\right) \times$$

$$f\left(\mathbf{X} \mid \eta_X\right) \times$$

$$f\left(\eta_M \mid \eta_X\right) \times$$

$$f\left(\eta_Y \mid \eta_X, \eta_M\right)$$

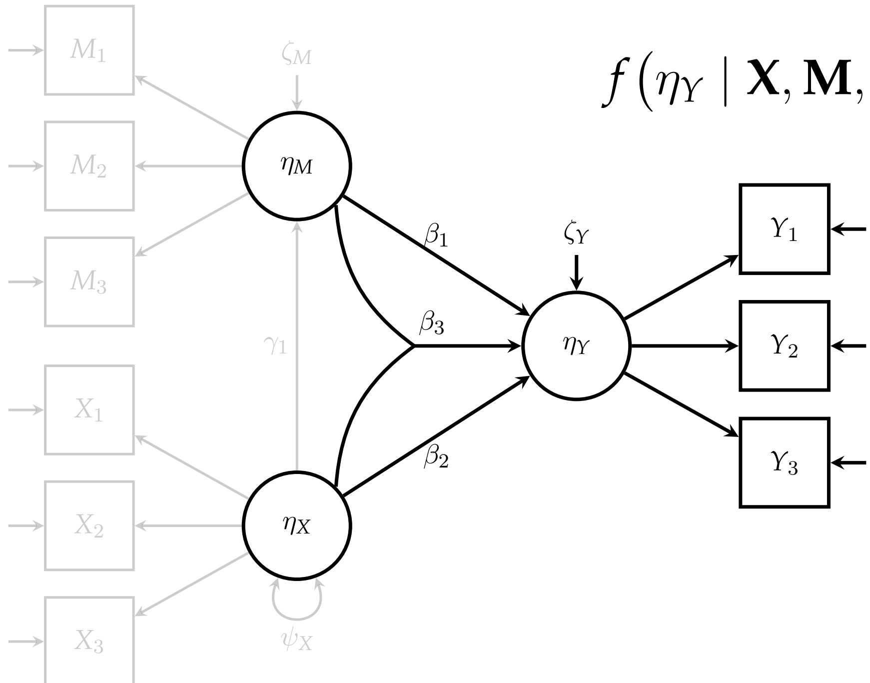


$$f\left(\eta_M \mid \mathbf{X}, \mathbf{M}, \mathbf{Y}, \eta_X, \eta_Y\right) \propto$$

$$f\left(\eta_M \mid \eta_X\right) \times$$

$$f\left(\mathbf{M} \mid \eta_M\right) \times$$

$$f\left(\eta_Y \mid \eta_X, \eta_M\right)$$



$$f\left(\eta_Y \mid \mathbf{X}, \mathbf{M}, \mathbf{Y}, \eta_X, \eta_M\right) \propto f\left(\mathbf{Y} \mid \eta_Y\right) \times f\left(\eta_Y \mid \eta_X, \eta_M\right)$$

Analyzing Multiply Imputed Factor Scores

An advantage of using the Bayesian-based approaches is obtaining factored scores for any latent factor.

These factor scores are a byproduct of the estimation procedure, where the latent variables are treated as missing variables that must be imputed.

1. Fit the model via Bayes

2. Save multiple copies of the latent factor scores

3. Fit an ordinary least squares regression for each copy

4. Use multiple imputation pooling formulas (Rubin, 1987) to pool the estimates and standard errors

Monte Carlo Simulation Study

$$\eta_Y = \alpha_Y + \beta_1 \eta_{Mi} + \beta_2 \eta_{Xi} + \beta_3 \eta_{Mi} \times \eta_{Xi} + \zeta_{Yi}$$

Coefficients	**% Bias**		**Coverage**		**RMSE Ratio**
	Imp	Bay	Imp	Bay	(Imp / Bay)
η_X (β_1)	1.043	0.801	0.932	0.934	1.009
η_M (β_2)	0.683	0.505	0.948	0.942	1.008
$\eta_X \times \eta_M$ (β_3)	1.319	1.005	0.966	0.956	1.012

Evaluating Latent Interactions in Blimp

Traditional SEM fit indices and tests generally do not perform well when testing for interactions (Asparouhov & Muthén, 2021).

This is due to likelihood-based SEM approaches fitting to only the first and second-order moments of the distribution (mean vector and covariance matrix).

Using the imputed factor scores allows us to simplify the problem and use traditional regression diagnostics.

Residual Based Diagnostics

Compute the residuals based on fitting a linear regression to each imputed data set.

Investigate classical regression diagnostics based on the residuals. Examples:

- Plotting Residuals versus Predicted scores
- Evaluating if correlation remains among residuals.

1. Generate imputations of latent factor scores and look at posterior mean of factor scores.

2. Fit a loess model to each imputed data set and obtain the fitted values and standard errors.

3. Pool across all analysis results and create a plot to investigate the fit.

Imputation Model 1:

$$\eta_Y = \alpha_Y + \beta_1\eta_{Mi} + \beta_2\eta_{Xi} + \beta_3(\eta_{Mi} \times \eta_{Xi}) + \zeta_{Yi}$$

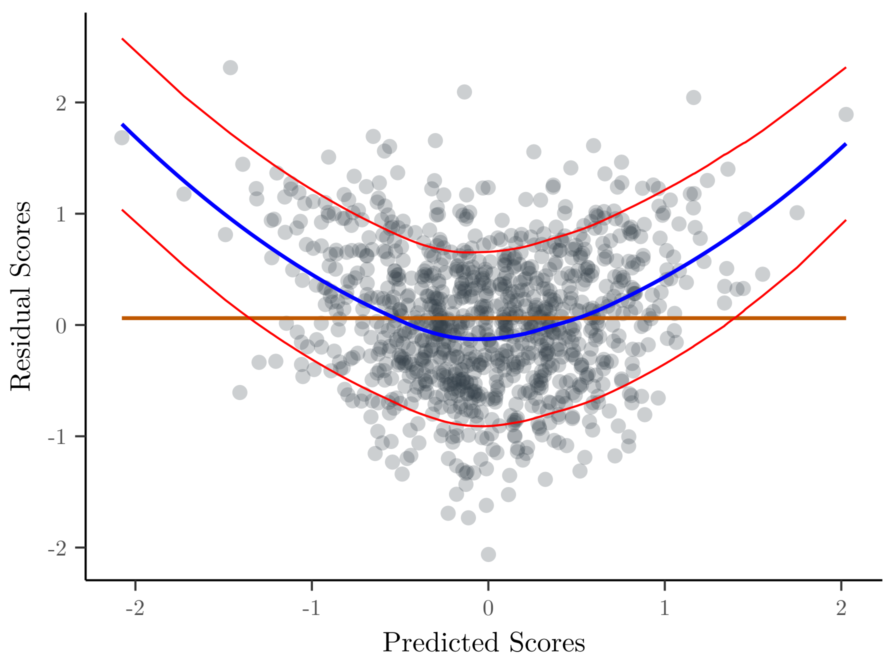
- Generate imputations based on **Model 1**
- Fit regression model without interaction
- Plot Residuals v. Predicted

Model 1 without Latent Interaction

Averaged over 100 imputations

Model 1 without Latent Interaction

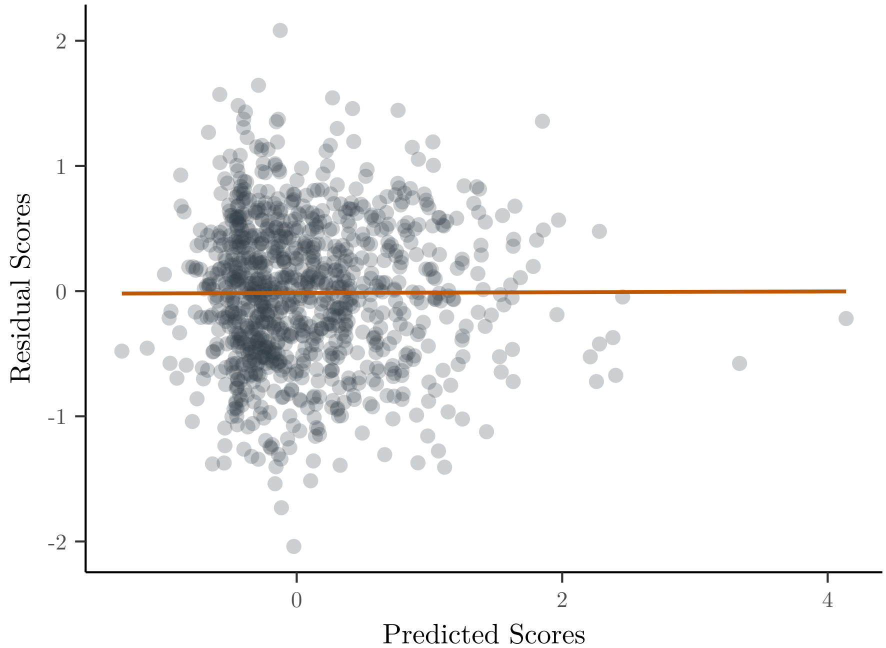
Averaged over 100 imputations



Brian T. Keller (University of Missouri) Latent Variable Models with Blimp @ Yonsei University (June 26, 2024) 53

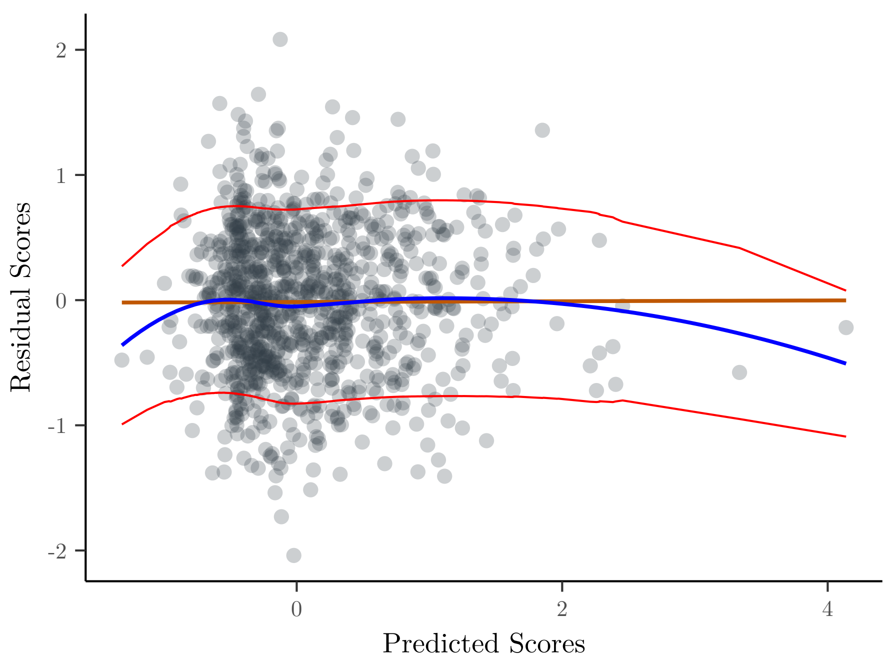
Model 1 with Latent Interaction

Averaged over 100 imputations



Model 1 with Latent Interaction

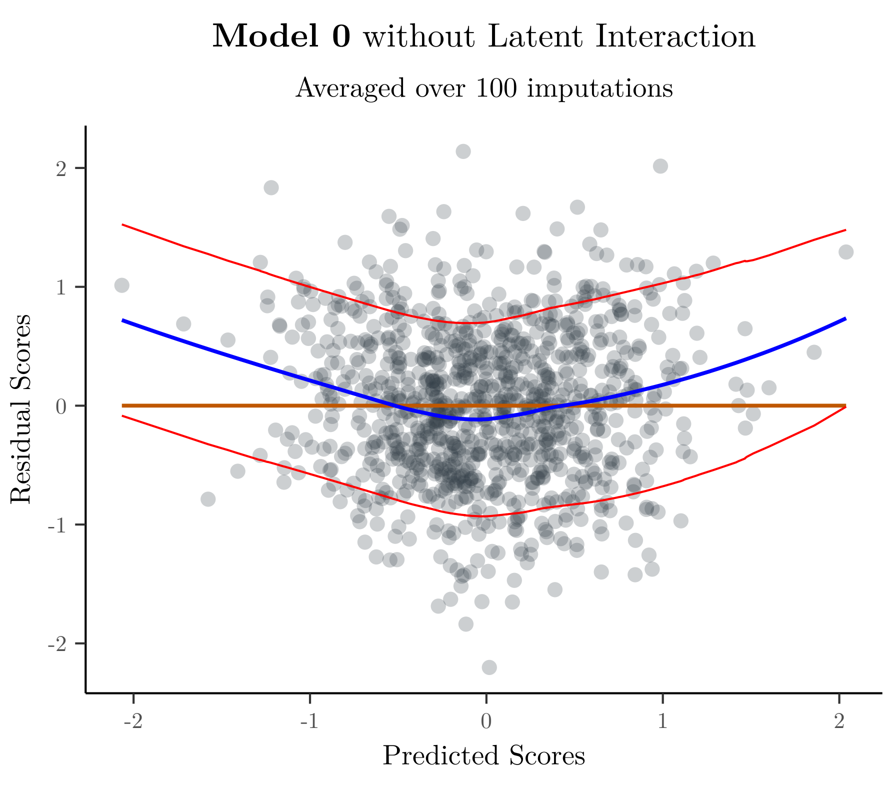
Averaged over 100 imputations

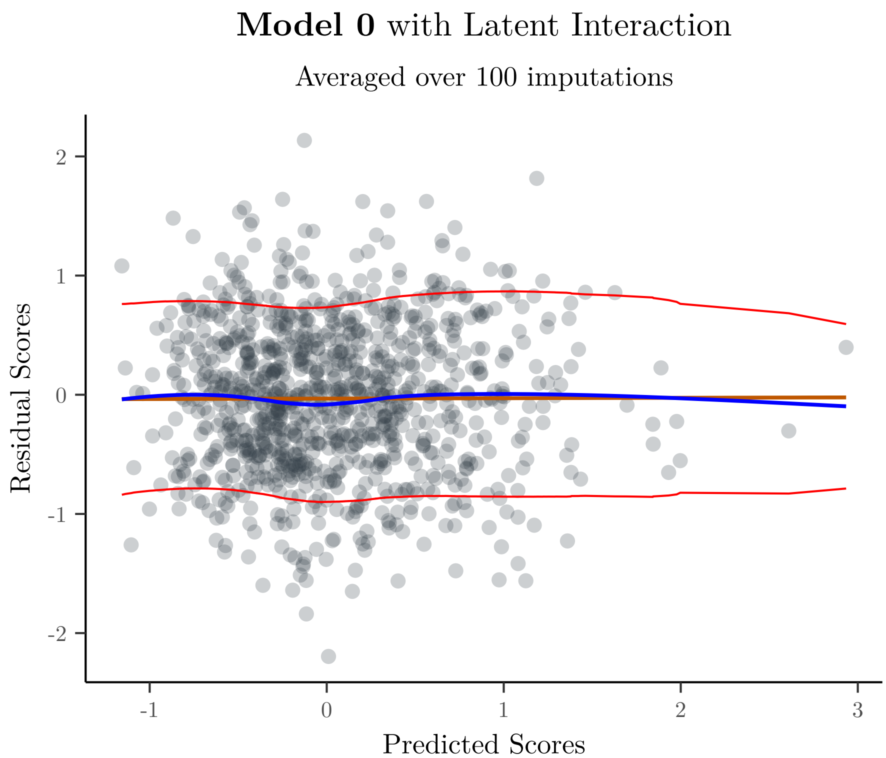


Suppose a researcher is not expecting a latent $\times$ latent interaction.

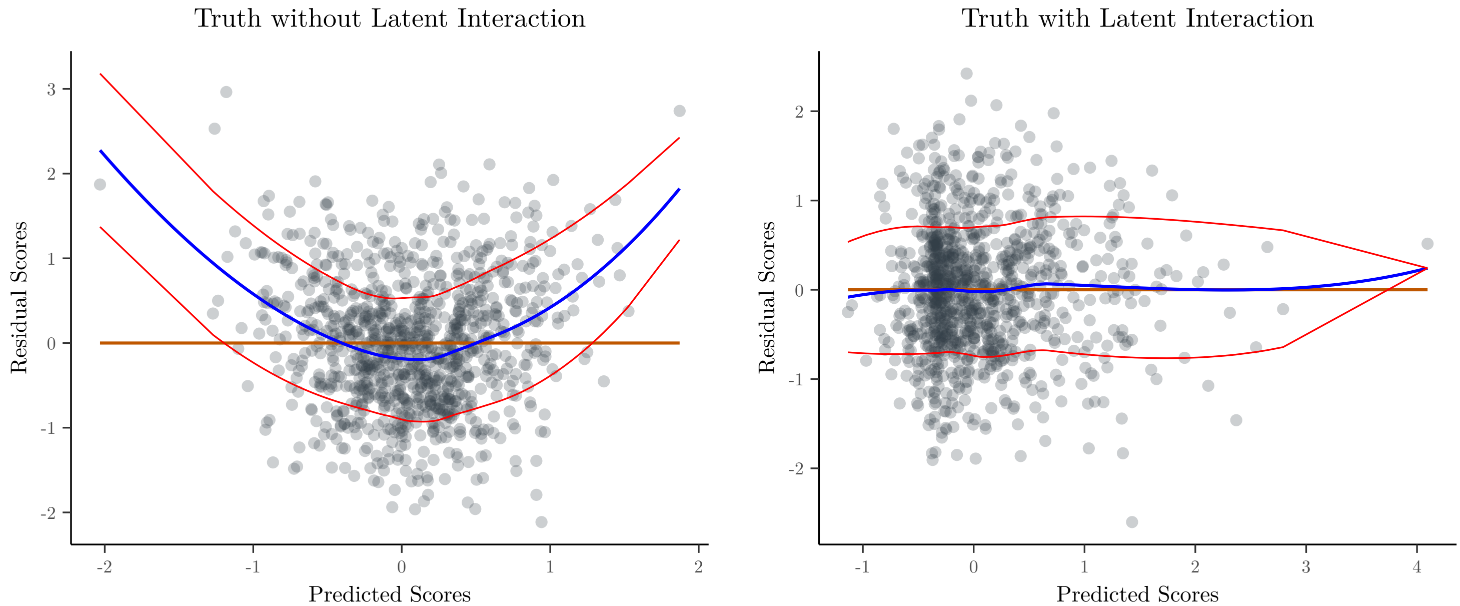
Imputation Model 0:

$$\eta_Y = \alpha_Y + \beta_1 \eta_{Mi} + \beta_2 \eta_{Xi} + \zeta_{Yi}$$

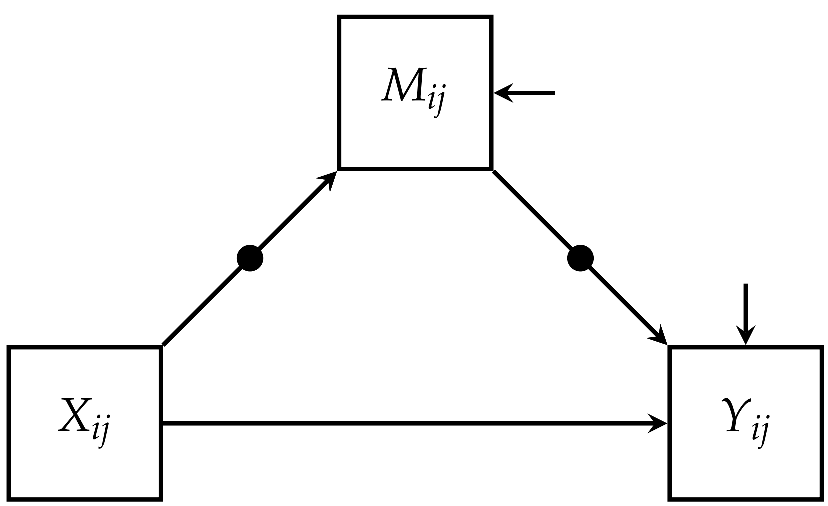




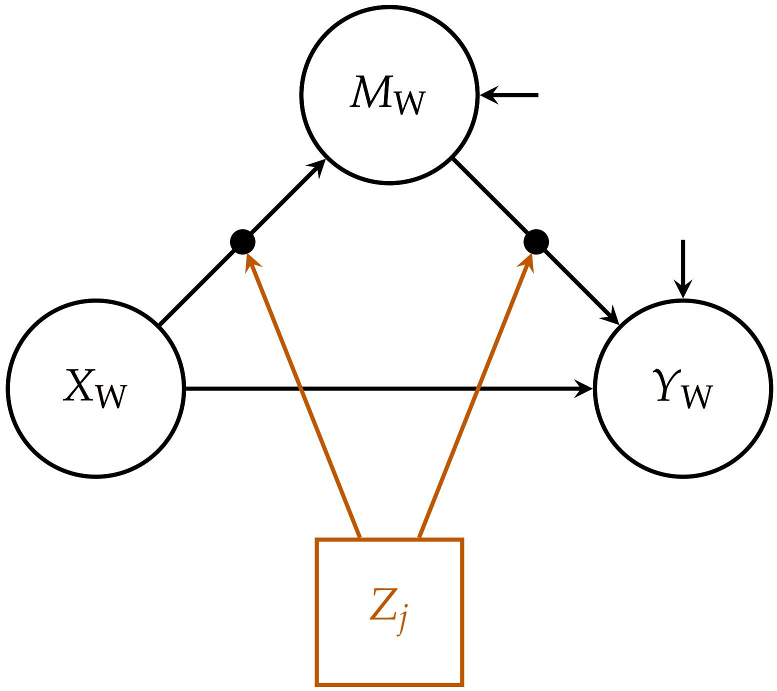
Plots of True Score Values



Multilevel Analysis Example in Blimp



$$Y_{ij} = Y_{\mathrm{W}} + Y_{\mathrm{B}}$$



$$
\begin{aligned}
\textbf{Level-1:} \quad M_{ij} &= M_{B} + a_j(X_{ij} - X_{B}) + r_{ij} \\
\textbf{Level-2:} \quad M_{B} &= \alpha_0 + \alpha_2(Z_j - \mu_Z) + u_{0_j} \\
a_j &= \alpha_1 + \alpha_3(Z_j - \mu_Z) + u_{1_j}
\end{aligned}
$$

$$
\begin{aligned}
\textbf{Level-1:} \quad Y_{ij} &= Y_{B} + b_j(M_{ij} - M_{B}) + \beta_2(X_{ij} - X_{B}) + e_{ij} \\
\textbf{Level-2:} \quad Y_{B} &= \beta_0 + \beta_3(Z_j - \mu_Z) + v_{0_j} \\
b_j &= \beta_1 + \beta_3(Z_j - \mu_Z) + v_{1_j}
\end{aligned}
$$

Blimp Syntax: Predictor Models

```
LATENT: # Declare Latent Variable Names
    id = y_b m_b x_b a_j b_j;

MODEL: # Begin Modeling syntax

    predictor_model:  # Label Block for organization
        # X within and between models
        x_ij ~ intercept@x_b;
        x_b ~ intercept;

        # Z Between model
        z_j ~ intercept@mu_z;
```


Blimp Syntax: M and Y Models

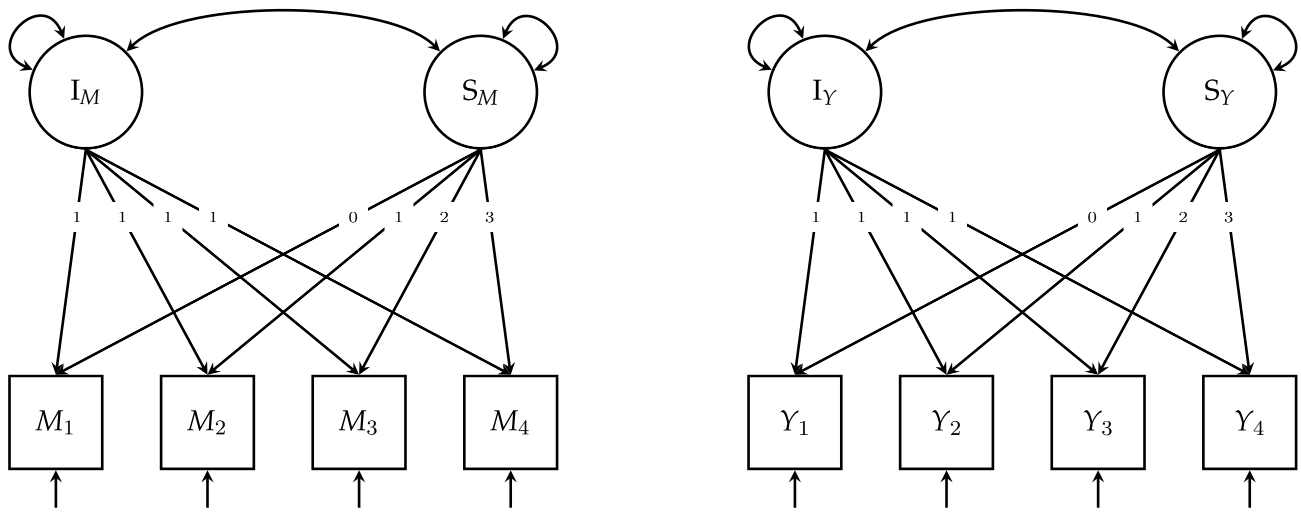
```
    # M within and between model
    mediator_model:
       m_ij ~ intercept@m_b  (x_ij - x_b)*a_j@1;
       m_b ~ intercept (z_j - &mu_z);
       a_j ~ intercept@alpha_1 (z_j - &mu_z)@alpha_3;

    # Y within and between model
    outcome_model:
       y_ij ~ intercept@y_b (m_ij - m_b)*b_j@1 (x_ij - x_b);
       y_b  ~ intercept (z_j -  &mu_z);
       b_j ~ intercept@beta_1;
```

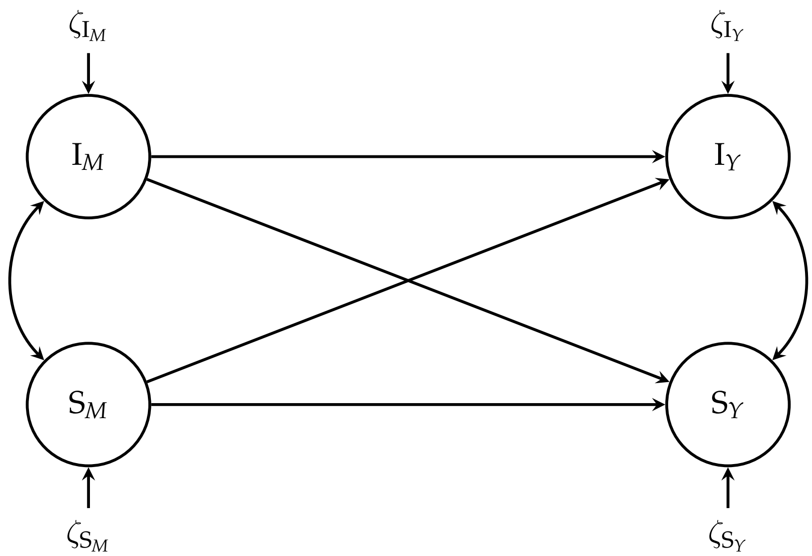
Syntax: Compute Indirect Effects

```
PARAMETERS: # Define Parameters

    # Define the Standard deviation of Z
    sd_z = sqrt(z_j.totalvar);

    ## Indirect effects for...
    # Low Z
    indirect.low = ((alpha_1 − (alpha_3 ∗ sd_z)) ∗ beta_1);
    # Mean Z
    indirect.med = alpha_1 ∗ beta_1;
    # High Z
    indirect.high = ((alpha_1 + (alpha_3 ∗ sd_z)) ∗ beta_1);
```


Parallel Growth Process Example in Blimp

We have two parallel growth processes:





Blimp Syntax: M Model

```
MODEL:
    m_model: # Label for organization
        I_m -> m.1@1 m.2@1 m.3@1 m.4@1;
        S_m -> m.1@0 m.2@1 m.3@2 m.4@3;
        m.1 ~ 0; m.2 ~ 0; m.3 ~ 0; m.4 ~ 0;

        m.1 ~~ m.1@var_m; # Fix variances to be equal
        m.2 ~~ m.2@var_m;
        m.3 ~~ m.3@var_m;
        m.4 ~~ m.4@var_m;

        I_m ~~ S_m;  # Correlation
```


Blimp Syntax: Y Model

```
    y_model:
        I_y -> y.1@1 y.2@1 y.3@1 y.4@1;
        S_y -> y.1@0 y.2@1 y.3@2 y.4@3;
        y.1 ~ 0; y.2 ~ 0; y.3 ~ 0; y.4 ~ 0;

        y.1 ~~ y.1@var_y; # Fix variances to be equal
        y.2 ~~ y.2@var_y;
        y.3 ~~ y.3@var_y;
        y.4 ~~ y.4@var_y;

        I_y ~~ S_y;  # Correlation
```

Blimp Syntax: Structural Model

```
structural_model:
    I_y ~ Intercept I_m S_m;

    S_y ~ Intercept I_m S_m;
```


What about latent by latent interactions?

Blimp Syntax: Structural Model Revised

```
structural_model:
    I_y ~ Intercept I_m S_m  # Main effects
            I_m*S_m;         # Two Way interaction

    S_y ~ Intercept I_m S_m  # Main effects
            I_m*S_m;         # Two Way interaction
```


What if I have a moderated mediation model with observed predictor?

Blimp Syntax: Structural Model Revised

```
structural_model:
    I_y ~ Intercept X I_m S_m   # Main effects
            X*I_m X*S_m I_m*S_m # Two Way interactions
            X*I_m*S_m;          # Three Way interaction

    S_y ~ Intercept X I_m S_m   # Main effects
            X*I_m X*S_m I_m*S_m # Two Way interactions
            X*I_m*S_m;          # Three Way interaction

    # M Models
    I_m ~ Intercept X;
    S_m ~ Intercept X;
```


Preliminary Results from Monte Carlo Study

Results for Latent Intercept of Y

	Prop. Bias	LCL (95%)	UCL (95%)
Intercept	−0.033	−0.065	−0.002
X	−0.029	−0.116	0.058
I_m	0.032	−0.010	0.073
S_m	0.009	−0.019	0.036
$\mathrm{I}_m \times X$	0.055	−0.045	0.154
$\mathrm{S}_m \times X$	−0.025	−0.089	0.040
$\mathrm{I}_m \times \mathrm{I}_m$	0.036	−0.006	0.079
$\mathrm{I}_m \times \mathrm{S}_m \times X$	−0.084	−0.242	0.074
Resid Var.	−0.003	−0.007	0.002

Results for Latent Slope of Y

	Prop. Bias	LCL (95%)	UCL (95%)
Intercept	-0.013	-0.044	0.018
X	-0.013	-0.101	0.075
I_m	0.024	-0.018	0.067
S_m	-0.003	-0.030	0.023
$\mathrm{I}_m \times X$	0.027	-0.070	0.124
$\mathrm{S}_m \times X$	-0.048	-0.111	0.015
$\mathrm{I}_m \times \mathrm{I}_m$	0.021	-0.020	0.062
$\mathrm{I}_m \times \mathrm{S}_m \times X$	0.014	-0.135	0.163
Resid Var.	-0.004	-0.008	0.000

Average ESS for Latent Intercept of Y

	Blimp	Mplus	Ratio
Intercept	10818	11140	0.97
X	7949	978	8.13
I_m	4449	4290	1.04
S_m	7500	7051	1.06
$\mathrm{I}_m \times X$	4738	557	8.51
$\mathrm{S}_m \times X$	6624	907	7.30
$\mathrm{I}_m \times \mathrm{I}_m$	4851	4503	1.08
$\mathrm{I}_m \times \mathrm{S}_m \times X$	5168	923	5.60
Resid Var.	8343	1481	5.63

Average ESS for Latent Slope of Y

	Blimp	Mplus	Ratio
Intercept	11705	11704	1.00
X	8151	972	8.38
I_m	4615	4423	1.04
S_m	7523	7286	1.03
$\mathrm{I}_m \times X$	4821	545	8.84
$\mathrm{S}_m \times X$	6649	897	7.41
$\mathrm{I}_m \times \mathrm{I}_m$	5075	4691	1.08
$\mathrm{I}_m \times \mathrm{S}_m \times X$	5281	907	5.82
Resid Var.	8789	1467	5.99

Approximate run time for one replication:

(Warm Up = 10000, Post = 20000, 2 chains)

- **Mplus** $\approx$ 12 minutes
- **Blimp** $\approx$ 36 seconds

Some Other Examples...

- Simple Slopes with Latent Interactions
- Multi-Group Models with Incomplete Groups
- Latent Moderated Mediation with Quadratic Relationship
- Latent Moderated Mediation with Skewed Indicators
- Multilevel Mediation with Ordinal Mediator and Latent Means

Blimp Software

- General-purpose Bayesian estimation for regression and path models.
- Single and multilevel models
- Allows for latent variables, incomplete predictors and outcomes
- Interactive and nonlinear effects
- Nonnormal data
- And more!

Available at

`https://www.appliedmissingdata.com/blimp`