Using Monte Carlo simulations to understand probabilities and modeling: bringing causality into the teaching of introductory statistical modeling

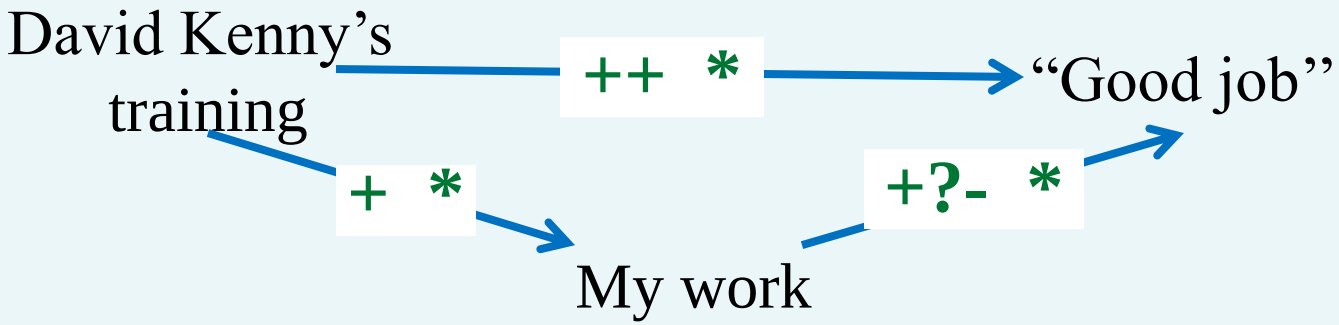
Applying Models to Stats

Emil Coman[1], Maria Coman[2], Eugen Iordache[3], Lisa Dierker[4], and Russell Barbour[5]

[1]U. of Connecticut Health Center, [2] Eastern Conn State U., [3] Transilvania U., Romania, [4] Wesleyan U., [5] Yale U.

Modern Modeling Methods Conference, Storrs, CT, May 21-22, 2013

Acknowledgment



DK: 'we all male mistakes'

Plan of attack

- Ease in using Monte Carlo data simulation in Excel and Mplus.
- Comparing simple causal model fit testing to MC models where the population model is the covariance matrix of the sample data.
- Interpreting statistical significance testing and model fit in regular model testing vs. MC models.

Number generation and uses

- One can use Excel to "gain insight into the workings of many procedures" by means of simulations of data.
- Monte Carlo simulations are used to answer *What if*'s, yet we show here how to use them to better understand the mechanics of SEM.
- Generating data based on summary statistics like means, variances, covariances (or regression coefficients) is *latent variable modeling* at its best.
- It is similar to generating plausible values for completely or partially missing data based on information from summary statistics and other related variables.

Miles, J. N. V. (2005). Confirmatory factor analysis using Microsoft Excel. *Behavior Research Methods, 37*(4), 672-676. doi: 10.3758/bf03192739

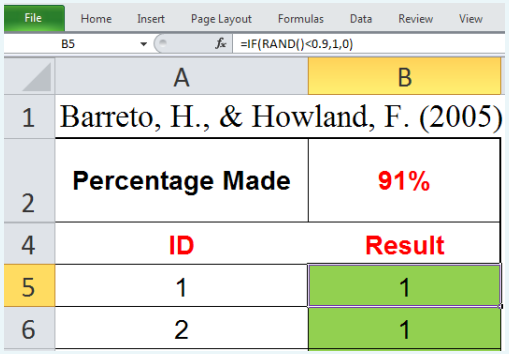
Random number generation

One can use Excel for an entire class in intro and intermediate and even advanced research methods course; yes, one can 'run' SEM in Excel too.

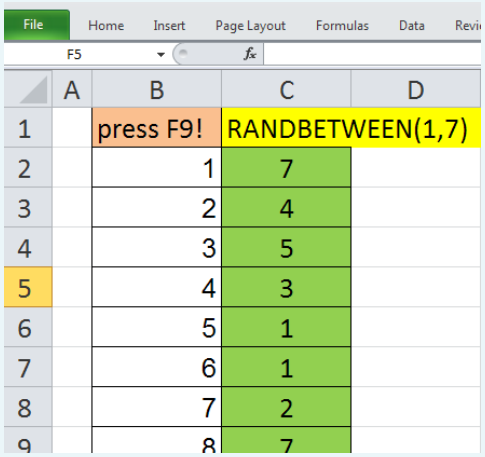
Generating 'truly' random variables:

[see file]

1. RAND() = continuous between 0 & 1



2. dichotomy; use 1. and split at some threshold, say .9.



3. RANDBETWEEN(1,7) integers between 1 & 7

Barreto, H., & Howland, F. (2005). *Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel: Cambridge University Press.*

Random numbers and inference

Excel can easily show how a sample mean becomes a variable, and its variability becomes the long misunderstood Standard Error:

- a mean → many means → mean of means → standard deviation of the many means imaginary variable

			0.61	0.98	0.07
			0.48	0.45	0.24
			0.92	0.08	0.26
			0.70	0.46	0.22
Mean of Means	0.50	Means	0.45	0.62	0.40
Mean of SDs	0.29	SDs	0.34	0.32	0.30
Mean of Variances		Variances	0.12	0.10	0.09

Randomness and DGP

'Random number' can mean: each new generation is independent of the previous one OR the numbers are not dependent on something else. The 2nd is often forgotten.

The Data Generating Process (DGP, Barreto & Howland 2005) can be

- fully random and causally blind, or
- accommodate reasonable causal mechanisms

In between there is DGP based on covariances (correlations): one can generate two variables X and Y that are correlated σ^2_{XY} (or ρ_{XY}).

Barreto, H., & Howland, F. (2005). *Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel: Cambridge University Press.*

Simple ways of generating data in Mplus

Generating a latent variable 'from nothing':

```
MODEL   :
LATENTX by;
LATENTX @1 ;
[LATENTX @0] ;
```

2nd option

```
MODEL   :
LATENTX by X@0;
LATENTX @1 ;
[LATENTX @0] ;
```

Another option of a new variable defined with 'old' ones:

```
!define LCS scores
dL21 by ; dL21@0;
dL21 on X1@-1 X2@1;
Linda Muthen
```

Mplus discussion forum

Muthén, L. K., & Muthén, B. O. (1998-2010). Mplus User's Guide. (Sixth ed.). Los Angeles, CA: Muthén & Muthén.

For more see

'1. Testing Mediation the Way it was Meant to be: Changes leading to changes then to other changes. Dynamic mediation implemented with latent change scores' by Emil Coman, Eugen Iordache, and Maria Coman; **Extensions to Mediational Analyses**

And posters:

2. 'Changes in Risk Behavior Achieved by Activating Dynamic Coupling Processes: dynamic growth modeling of a health prevention intervention' by Emil Coman, Carolyn Lin, Suzanne Suggs, Eugen Iordache, Maria Coman, and Russell Barbour &

3. Investigating the Directionality and Pattern of Mutual Changes of Health Outcomes: Adding dynamic perspectives to static longitudinal analyses by Emil Coman, Marco Bardus, Suzanne Suggs, Eugen Iordache, Maria Coman, and Holly Blake

Simple ways of generating data in Mplus

Generating a variable using known mean and variance:

Generated data can be saved, and one can compute mean and variance and their standard error, and compare them to the MC generated values.
Output of the simplest MC 'study':

```
MODEL
POPULATION:
scon1@0.575 ;
[scon1@3.539] ;
MODEL   :
scon1*0.575 ;
[scon1*3.539] ;
```

MODEL RESULTS

		ESTIMATES		S. E.	M. S. E.	95%	% Sig
	Population	Average	Std. Dev.	Average		Cover	Coeff
Means							
SCON1	3.539	3.5385	0.0410	0.0416	0.0017	0.960	1.000
SPSS	*3.589*		*.044*$_{SE\mu}$				
Variances							
SCON1	0.575	0.5746	0.0453	0.0447	0.0020	0.936	1.000
	.645						

Parameter estimates are obtained "over the repeated draws of independent samples, referred to as replications" (Muthen, 2002).
In italics are the SPSS numbers from 1 generated sample. The SEμ across imaginary resamplings becomes in MC the 'Std. Dev.' across replications.

Muthén, B. (2002). Using Mplus Monte Carlo simulations in practice: A note on assessing estimation quality and power in latent variable models. *Mplus Web Notes, 1(2).*

Simple ways of generating data in Mplus

MODEL RESULTS

		ESTIMATES		S. E.	M. S. E.	95%	% Sig
	Population	Average	Std. Dev.	Average		Cover	Coeff
Means							
SCON1	3.539	3.5385	0.0410	0.0416	0.0017	0.960	1.000
SPSS	*3.589*		*.044* $_{SE\mu}$				
Variances							
SCON1	0.575	0.5746	0.0453	0.0447	0.0020	0.936	1.000
	.645						

Parameter bias for the mean is **100*(3.5385-3.539)/3.539 = - 0.014 = 1.4%**

Parameter bias for the variance is **100*(0.5746-0.575)/0.575 = - 0.069 = 6.9%**

1.4% more replications than the 5% expected by chance failed to find the population variance value within the 95% CI for the estimate.

The column labeled 95% Cover gives the proportion of replications for which the 95% confidence interval contains the population parameter value. This gives the coverage which indicates how well the parameters and their standard errors are estimated. In this output, the mean coverage value is close to the correct value of 0.95, while the variance coverage is slightly lower.

Two variable MC model

1st step
1. Run a model test

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
MOT2	ON				
ATT2		0.971	0.076	12.790	0.000

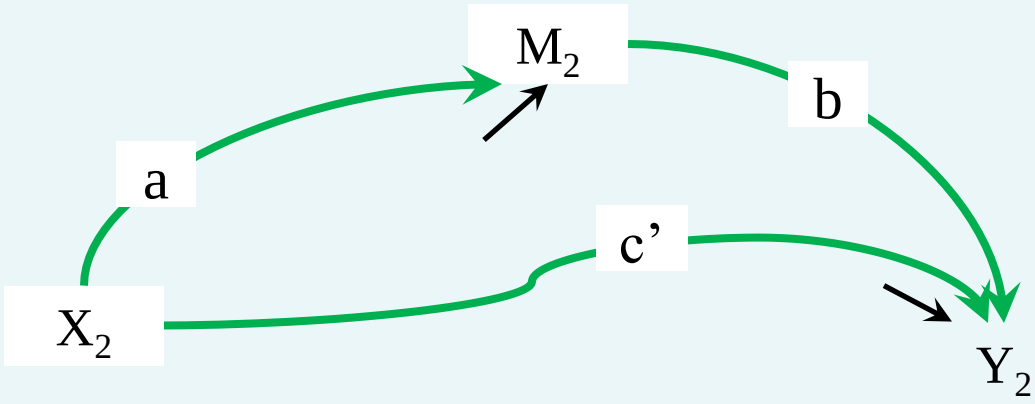
---

2. Run a 'MC covariance model' on sample means+covariances data

		ESTIMATES		S. E.	M. S. E.	95%	% Sig
	Population	Average	Std. Dev.	Average		Cover	Coeff
MOT2	ON						
ATT2	0.971	0.9676	0.0756	0.0761	0.0057	0.950	1.000

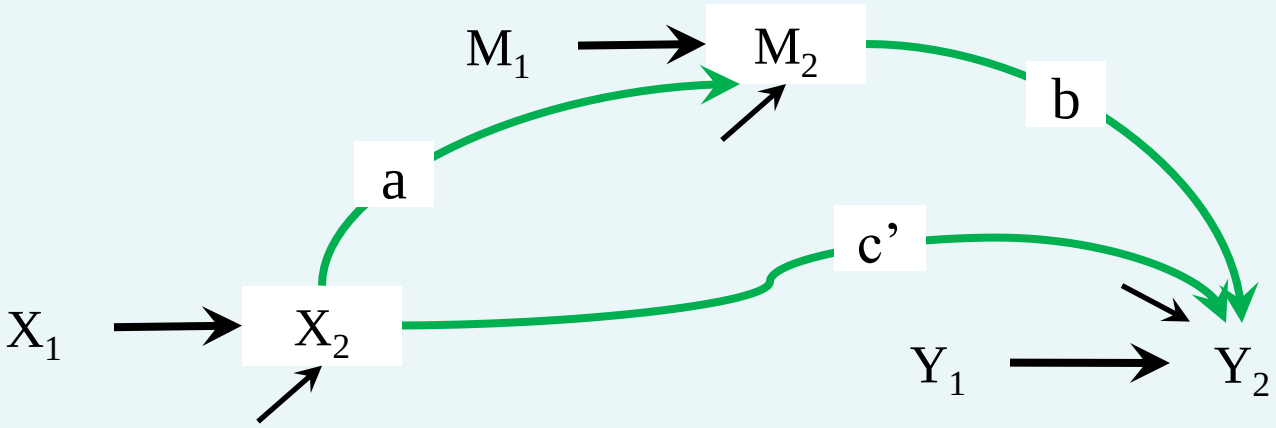
5% replications (as expected by chance) failed to find the population variance value within the 95% CI for the estimate.

Concurrent Barron-Kenny (no measurement errors)



Barron-Kenny mediation

Concurrent Barron-Kenny AR1 (no measurement errors)



Barron-Kenny mediation

Three variable BK mediation MC model

1. Model test

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SEF3 ON				
MOT2	0.606	0.082	7.409	0.000
MOT2 ON				
ATT2	0.971	0.076	12.790	0.000
Effects from ATT2 to SEF3				
Sum of indirect	0.588	0.092	6.411	0.000

χ^2 (1) = 1.180, p = .2773

---

2. MC model

		ESTIMATES		S. E.	M. S. E.	95%	% Sig
	Population	Average	Std. Dev.	Average		Cover	Coeff
SEF3 ON							
MOT2	0.606	0.6114	0.0657	0.0621	0.0043	0.939	1.000
MOT2 ON							
ATT2	0.971	0.9699	0.0749	0.0762	0.0056	0.959	1.000
Effects from ATT2 to SEF3							
Sum indirect	0.588	0.5931	0.0792	0.0763	0.0063	0.938	1.000

mean[χ^2 (1)] = 2.950, SD(χ^2) = 3.253

Proportions		Percentiles	
Expected	Observed	Expected	Observed
0.950	0.981	0.004	0.023
0.050	0.289	3.841	9.181

MC model fit nuances

χ^2 (1) = 1.180, p = .2773
Yet
mean[χ^2 (1)] = 2.950, SD(χ^2) = 3.253

Proportions		Percentiles	
Expected	Observed	Expected	Observed
0.050	0.289	3.841	9.181

28.9% of the simulations exceeded the 3.841 critical value (in excess of the 5% expected by chance variability alone).

If we re-sample from the same populations (not re-run the model!) 1,000 times, 29% models tested would not fit well, according to a .05 threshold.

-

Three variable AR mediation MC model vs. initial

Model test

In *blue italics* are the estimates from the mediation model with AR paths from same-prior variables

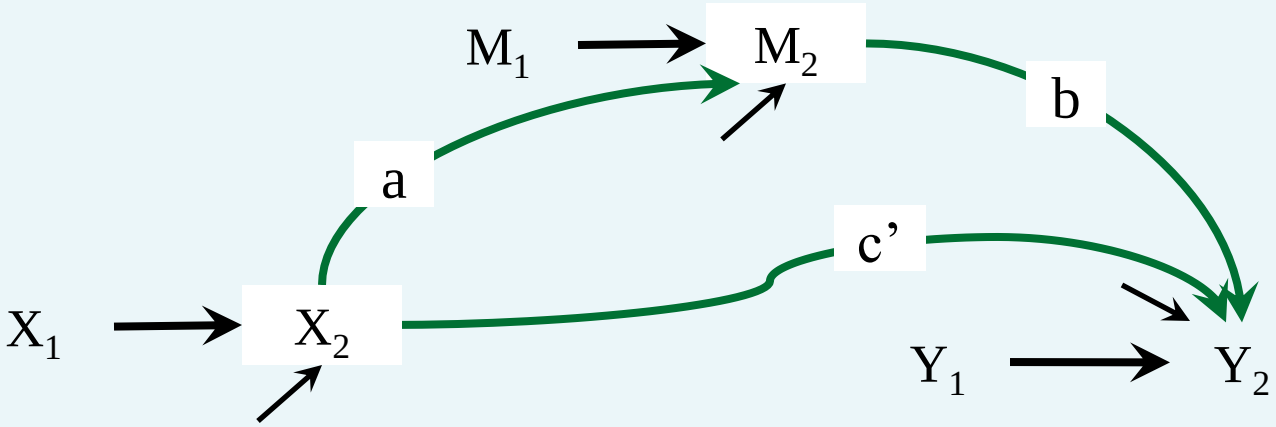
	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SEF3 ON				
MOT2	0.606	0.082	7.409	0.000
SEF3 ON				
MOT2	*0.486*	*0.086*	*5.673*	*0.000*
MOT2 ON				
ATT2	0.971	0.076	12.790	0.000
MOT2 ON				
ATT2	*0.854*	*0.075*	*11.348*	*0.000*
Effects from ATT2 to SEF3				
Sum of indirect	0.588	0.092	6.411	0.000
Effects from ATT2 to SEF3				
Sum of indirect	*0.416*	*0.082*	*5.086*	*0.000*

χ^2 (1) = 1.180, p = .2773

χ^2 (7) = 17.577, p =0.0140

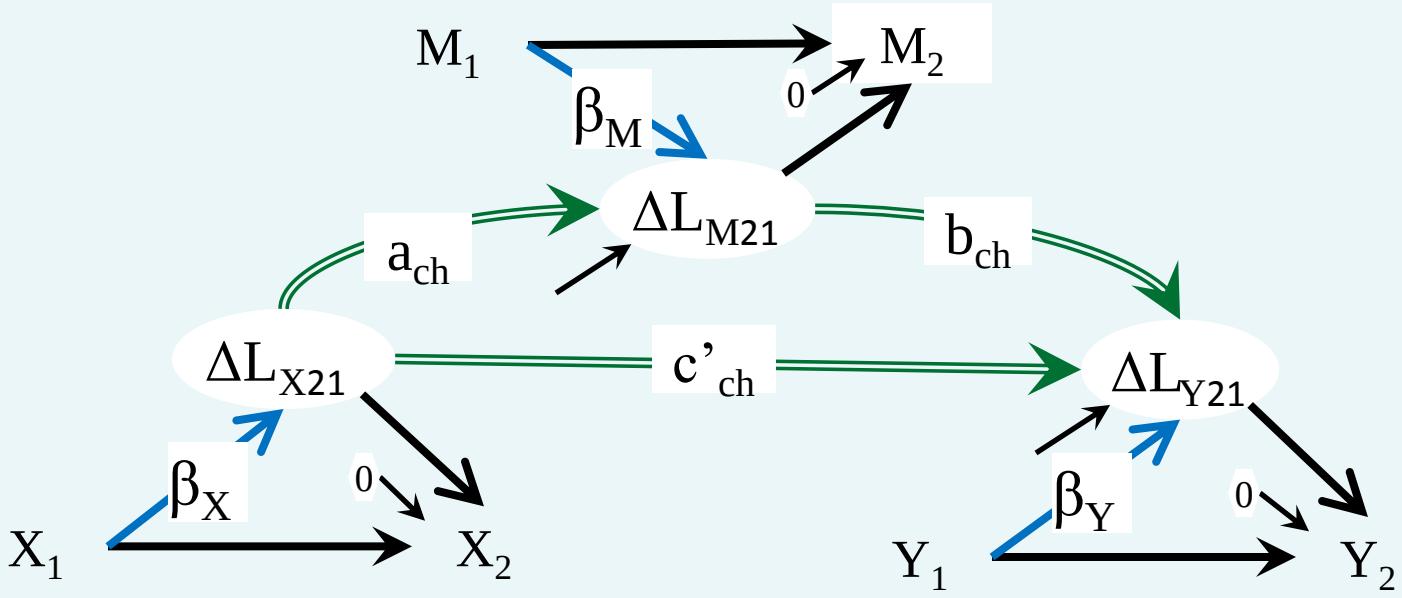
Lack of acceptable fit of the AR model indicates the need to model additional cross-variable mechanism, including the level-to-changes and changes-to-changes links.

Concurrent Barron-Kenny vs. Dynamic Mediation model (no measurement errors)

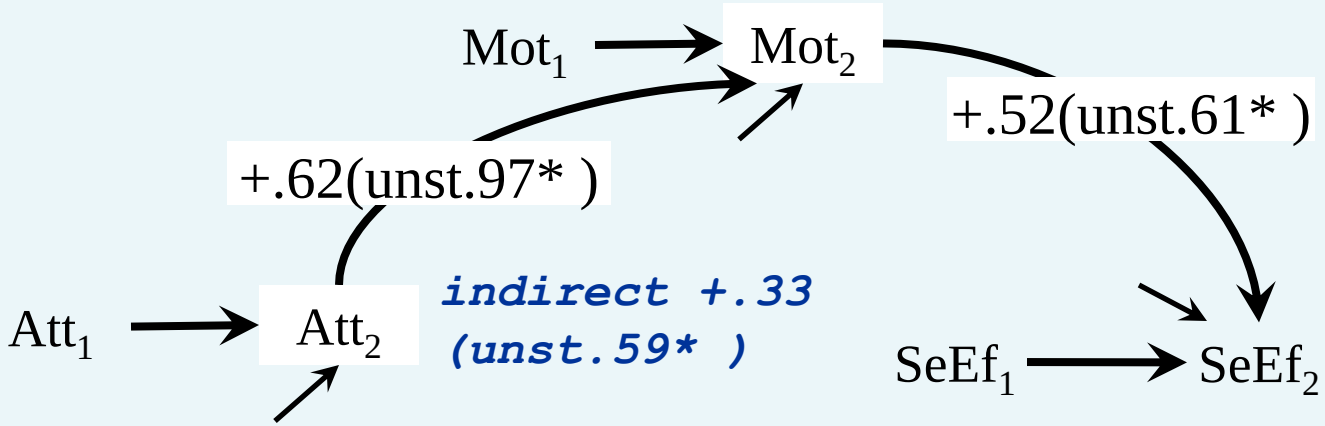


Barron-Kenny mediation

Dynamic mediation

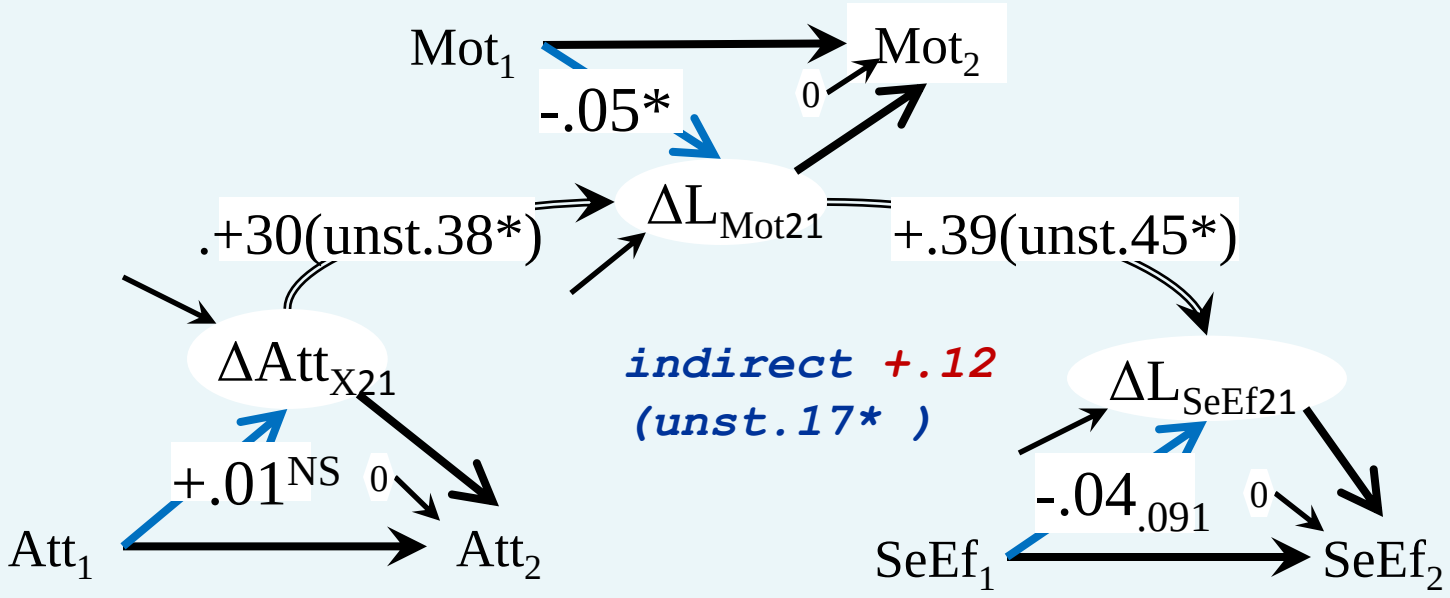


Concurrent Barron-Kenny vs. Physical Activity (PA) Dynamic Mediation
PA Attitude->PA Motivation->PA Self Efficacy



Barron-Kenny mediation

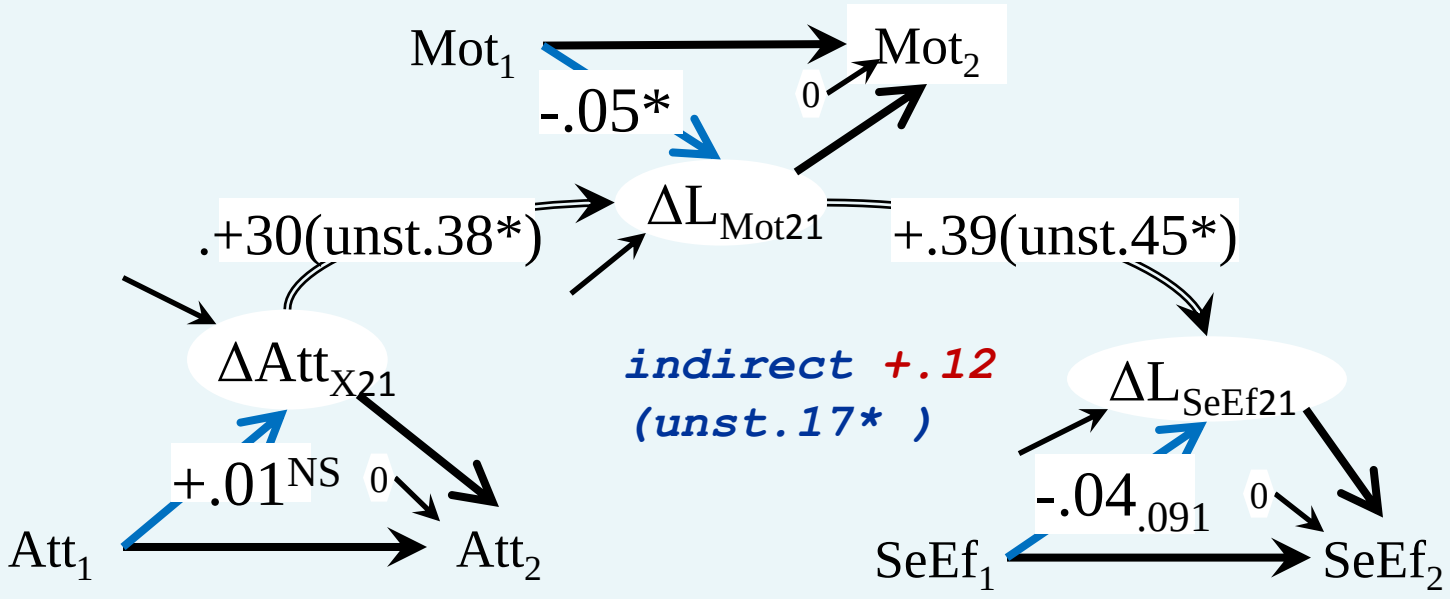
Dynamic mediation



LCS model fit unreportable.

Concurrent Barron-Kenny vs. Physical Activity (PA) Dynamic Mediation
PA Attitude->PA Motivation->PA Self Efficacy

Dynamic mediation



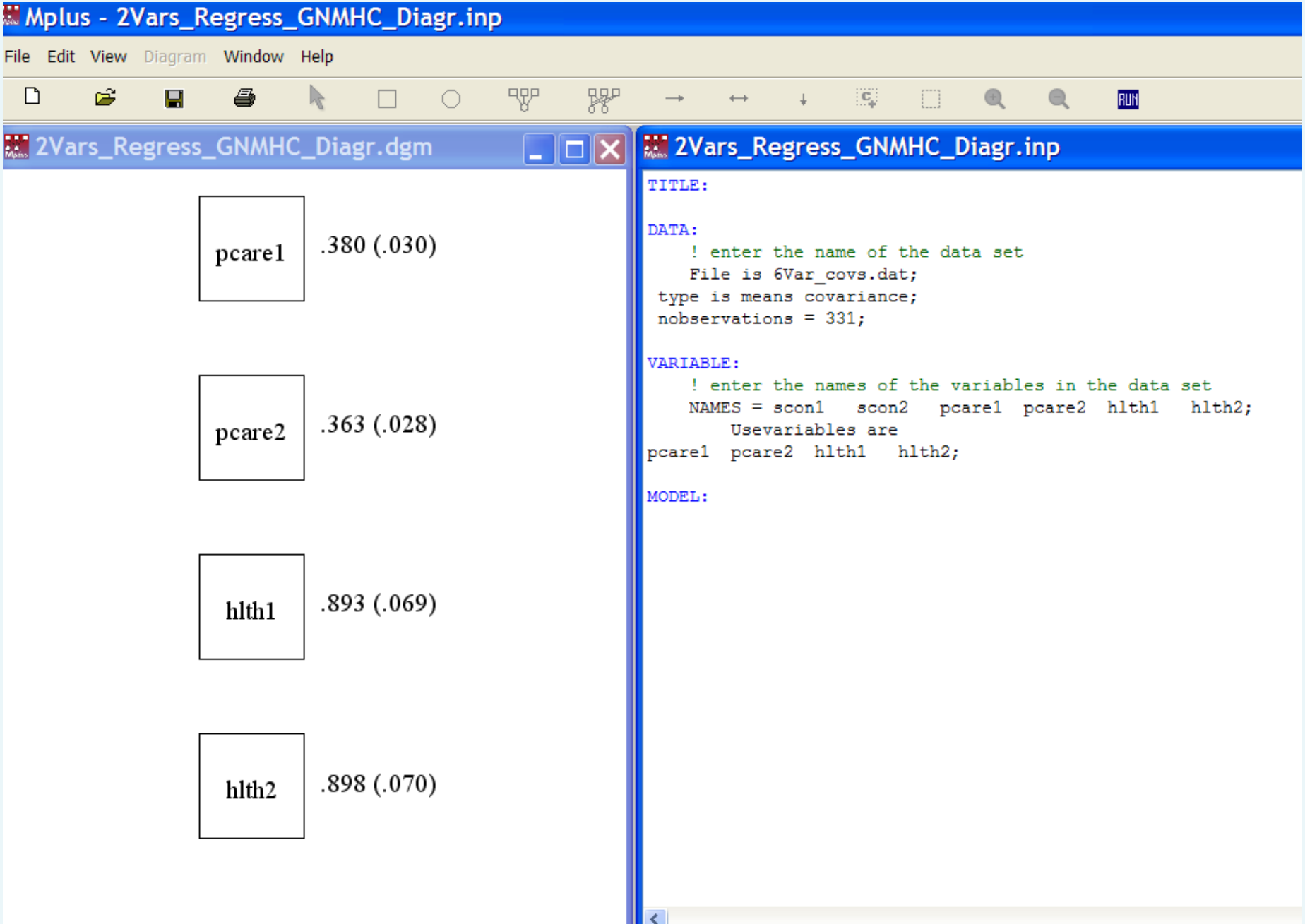
MC Model test

	ESTIMATES		S. E.	M. S. E.	95%	% Sig
Population	Average	Std. Dev.	Average		Cover	Coeff
BK Mediation with AR1 paths						
Effects from ATT2 to SEF3						
Sum indirect 0.588	0.5931	0.0792	0.0763	0.0063	0.938	1.000
Latent Change Score Mediation[A]						
Effects from DATT21 to DSEF21						
Sum indirect 0.174	0.2239	0.0426	0.0408	0.0043	0.791	1.000

Many more replications will fail to estimated the indirect effect within the 95% CI around the (assumed) population value of .174.

A: LCS model fit unreportable.

Mplus diagrammer Getting sample estimates

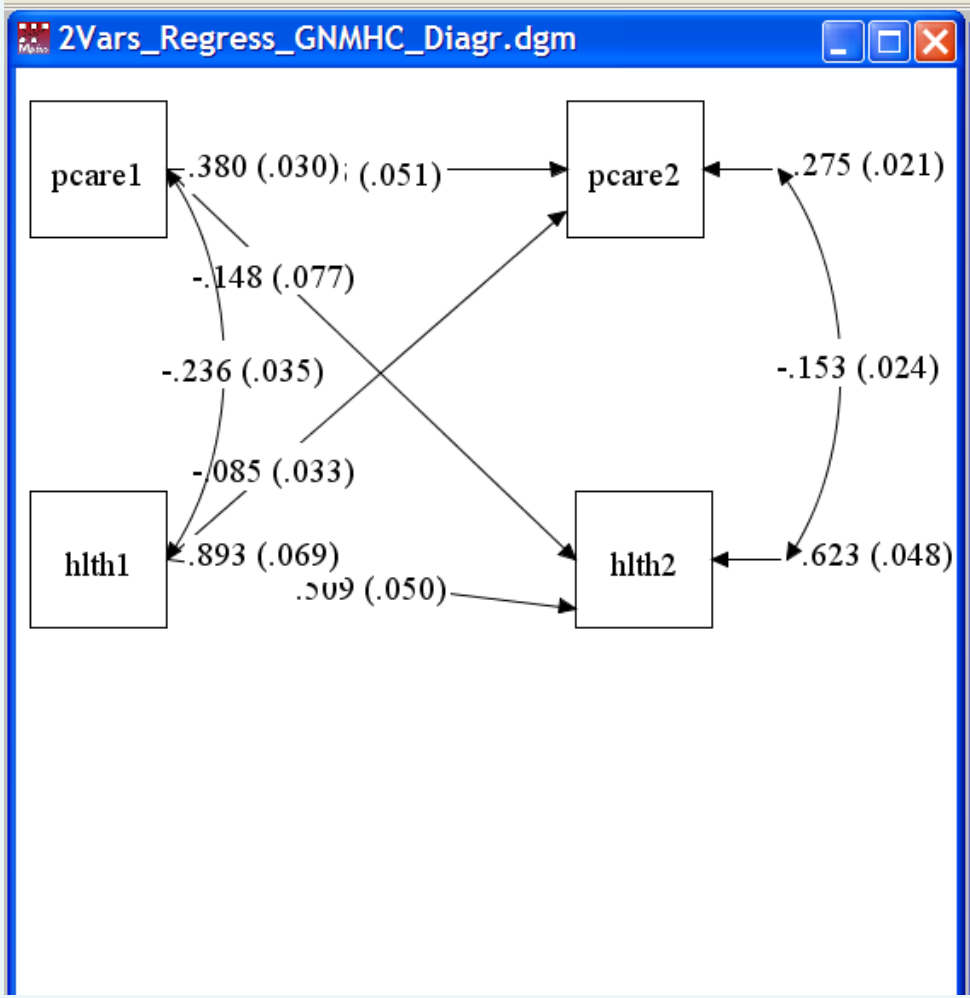


Mplus diagrammer +++

-

Mplus - 2Vars_Regress_GNMHC_Diagr.dgm

File Edit View Diagram Window Help



2Vars_Regress_GNMHC_Diagr.out

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
HLTH2 ON				
HLTH1	0.509	0.050	10.132	0.000
PCARE1	-0.148	0.077	-1.922	0.055
PCARE2 ON				
HLTH1	-0.085	0.033	-2.535	0.011
PCARE1	0.415	0.051	8.108	0.000
HLTH1 WITH				
PCARE1	-0.236	0.035	-6.839	0.000
HLTH2 WITH				
PCARE2	-0.153	0.024	-6.298	0.000
Means				
PCARE1	3.359	0.034	99.156	0.000
HLTH1	2.630	0.052	50.626	0.000
Intercepts				
PCARE2	2.368	0.224	10.563	0.000
HLTH2	1.719	0.338	5.092	0.000
Variances				
PCARE1	0.380	0.030	12.865	0.000
HLTH1	0.893	0.069	12.865	0.000
Residual Variances				
PCARE2	0.275	0.021	12.865	0.000
HLTH2	0.623	0.048	12.865	0.000

Conclusions

1. Monte Carlo simulations can be used to present basic concepts like standard errors, and statistical significance.
2. MC models make explicit the basis of the modeling assumptions behind models as simple as 1 variable- 1 group, but more complex one too.
   - Specifically they uncover the non-causal assumptions of the Data Generating Processes (DGP) believed to be at work in common examples used in teaching statistics (e.g. ‘2 random variables’ in regression).

Conclusions 2

3. MC models reposition both the model fit testing, and the parameter significance testing.
   - Well fitting models can yield in fact more than the expected by chance χ^2 above the threshold values (e.g. 3.841 for χ^2 (1).
4. The MC setup forces researchers to look back and find both trusted causal mechanism and reasonable values for population parameters they will focus on.

-

1 recent extension example

Method 2: Monte Carlo confidence intervals (MCCIs)

For example, if we want to obtain a 95% MCCI for the difference between two slopes b_1 and b_2, we could:

1. Estimate a model containing b_1 and b_2 as parameters.
2. Locate and record the estimates of b_1 and b_2.
3. Find the asymptotic variances and covariance of b_1 and b_2.
4. Generate thousands of pairs of b_1 and b_2 values; compute $(b_2 - b_1)$ for each pair.
5. Find the values cutting off the bottom and top 2.5% of the distribution of $(b_2 - b_1)$. These are the lower and upper bounds of a 95% MCCI.

Challenges and extensions

Questions:

- Using the raw *sample* covariance matrix+means vs. the model *estimated* covariance matrix+means
- Setting initial values – varying the range

✓Meta-analysis: one can use different summary parameters reported in repeated studies to assemble an aggregated/true population comprehensive causal DGP : this can be used to re-estimate effects from original studies under a broader specification of the models, than in each reported study.

(Coman & Huedo-Medina, CIRA-CHIP, unfunded proposal)

Thanks!

Questions?

coman@uchc.edu

-