A simulated data set for path analysis : Chi Square versus fit indices

By Alexandre J.S. Morin, October 6, 2008.

Hi all,

About 1-2 weeks ago John conducted a small simulation study and claimed based on its results that chi square (CS) was more sensitive to deviations from the true population model than the other fit indices (OFI). I disagreed (not on the legitimacy of the study) that the claims John made were supported by the results he got. So I agreed to pursue this further and to get back to SEMNET. This is what I am doing. The first part of my agreement was to start from John's path analysis simulated data set. The second part was to generate another simulated data set including latent variables. I am now reporting on the first part. I'll need another week or so to find time to work on the other one.

So, the population model for the simulated data set (n = 10 000) is (you should draw it, it is easier to grasp):

x1 = 10 + e1
x2 = 10 + x1 + e2
x3 = 10 + x2 + e3
x4 = 10 + x3 + e4
r = 20 + .3*x1 + e5
q = 15 + x2 + x3 + x4 + r + 2*e6
z = 10 + x3 + x4 + r + 2*e7
y = z + q + 2*r + e8

Note of Caution: None of this is a "real" simulation study in that we are working on a single artificial data set with no replications. So, some of the "weirder" results could be related to this fact (the data set generated from the population model could be an "extreme one" in some regards). But if I had done a real simulation study, I would probably attempt to publish it rather than posting it here. This is simply for debating.

Part 1. Alternative real models

Interestingly, the first "real" model John reported on was an incomplete one with the X1->X2->X3->X4 regressions taken out. Stan was the first to point this out. The fit of this model was equivalently good according to both the CS and the OFI. However, this model was still wrong in assuming that all Xs were exogenous when in fact only X1 was truly exogenous. But, since exogenous variables are by default allowed to correlate, the correlations that were included by default between X1-X2-X3-X4 did replace the regressions paths between them. So this model is, in the end, proper, even if it does not fully reflect the reality.

Full model:

Chi-Square (df) 7.970 (14), p = 0.8909

CFI 1.000 / TLI 1.000 /

RMSEA Estimate 0.000 /RMSEA 90 Percent C.I. 0.000 0.004 / SRMR 0.011

MODEL RESULTS

X2	ON	X1	1.017	0.010	101.424	0.000
X3	ON	X2	0.987	0.007	139.699	0.000
X4	ON	X3	1.000	0.006	169.898	0.000
R	ON	X1	0.303	0.010	29.875	0.000
Q	ON					
	X2		1.000	0.025	40.532	0.000
	X3		1.011	0.028	35.494	0.000
	X4		0.997	0.020	49.926	0.000
	R		0.998	0.020	50.784	0.000
Z	ON					
	X3		0.983	0.023	42.886	0.000
	X4		0.991	0.020	50.369	0.000
	R		0.980	0.019	50.971	0.000
Y	ON					
	Z		0.997	0.004	251.418	0.000
	Q		1.002	0.003	321.814	0.000
	R		2.012	0.010	198.902	0.000
Residual Variances						
	X2		0.989	0.014	70.711	0.000
	X3		1.001	0.014	70.711	0.000
	X4		1.024	0.014	70.711	0.000
	R		1.011	0.014	70.711	0.000
	Q		4.080	0.058	70.711	0.000
	Z		3.963	0.056	70.711	0.000
	Y		0.971	0.014	70.711	0.000

STANDARDIZED STDYX MODEL RESULTS

X2	ON	X1	0.712	0.005	144.460	0.000
X3	ON	X2	0.813	0.003	240.005	0.000
X4	ON	X3	0.862	0.003	334.940	0.000
R	ON	X1	0.286	0.009	31.179	0.000
Q	ON					
X2			0.259	0.006	40.368	0.000
X3			0.318	0.009	35.599	0.000
X4			0.364	0.007	50.007	0.000
R			0.192	0.004	47.363	0.000
Z	ON					
X3			0.391	0.009	43.370	0.000
X4			0.457	0.009	51.284	0.000
R			0.238	0.005	48.480	0.000
Y	ON					
Z			0.417	0.002	188.800	0.000
Q			0.529	0.002	235.545	0.000
R			0.204	0.002	106.856	0.000
Residual Variances						
X2			0.493	0.007	70.216	0.000
X3			0.339	0.006	61.490	0.000
X4			0.257	0.004	58.018	0.000
R			0.918	0.005	174.674	0.000
Q			0.137	0.003	54.332	0.000
Z			0.212	0.004	57.200	0.000
Y			0.009	0.000	51.952	0.000

<u>Alternative model</u> (I requested the exogenous correlations as they are not provided as default – this does not change the results):

TESTS OF MODEL FIT

Chi-Square (df) 6.236 (11), p = 0.8572

CFI 1.000 / TLI 1.000

RMSEA 0.000 / RMSEA 90 Percent C.I. 0.000 0.006 / SRMR 0.012

MODEL RESULTS

R	ON	X1	0.303	0.010	29.875	0.000
Q	ON					
	X2		1.000	0.025	40.532	0.000
	X3		1.011	0.028	35.494	0.000
	X4		0.997	0.020	49.926	0.000
	R		0.998	0.020	50.784	0.000
Z	ON					
	X3		0.983	0.023	42.886	0.000
	X4		0.991	0.020	50.369	0.000
	R		0.980	0.019	50.971	0.000
Y	ON					
	Z		0.997	0.004	251.418	0.000
	Q		1.002	0.003	321.814	0.000
	R		2.012	0.010	198.902	0.000
X1	WITH					
	X2		1.000	0.017	58.005	0.000
	X3		0.994	0.020	50.372	0.000
	X4		0.991	0.022	44.798	0.000
X2	WITH					
	X3		1.980	0.031	63.089	0.000
	X4		1.973	0.034	57.249	0.000
X3	WITH					
	X4		2.956	0.045	65.282	0.000
Variances						
	X1		0.984	0.014	70.711	0.000
	X2		2.006	0.028	70.711	0.000
	X3		2.956	0.042	70.711	0.000
	X4		3.980	0.056	70.711	0.000
Residual Variances						
	R		1.011	0.014	70.711	0.000
	Q		4.080	0.058	70.711	0.000
	Z		3.963	0.056	70.711	0.000
	Y		0.971	0.014	70.711	0.000

STANDARDIZED STDYX MODEL RESULTS

R	ON	X1	0.286	0.009	31.179	0.000
Q	ON					
X2		0.259	0.006	40.347	0.000
X3		0.318	0.009	35.608	0.000
X4		0.364	0.007	49.981	0.000
R		0.192	0.004	47.364	0.000
Z	ON					
X3		0.391	0.009	43.369	0.000
X4		0.457	0.009	51.275	0.000
R		0.238	0.005	48.484	0.000
Y	ON					
Z		0.417	0.002	188.643	0.000
Q		0.529	0.002	235.166	0.000
R		0.204	0.002	106.758	0.000
X1	WITH					
X2		0.712	0.005	144.460	0.000
X3		0.583	0.007	88.351	0.000
X4		0.501	0.007	66.905	0.000
X2	WITH					
X3		0.813	0.003	240.005	0.000
X4		0.698	0.005	136.253	0.000
X3	WITH					
X4		0.862	0.003	334.941	0.000
Residual Variances						
R		0.918	0.005	174.674	0.000
Q		0.137	0.003	54.290	0.000
Z		0.212	0.004	57.166	0.000
Y		0.009	0.000	51.889	0.000

Part 2. The absence of important variables ?

John also estimated "wrong" models based on a reduced data set and claimed that the CS better detect misspecifications due to the exclusion of important variables that are part of the population model. This is where I did react the most. Why?

Because with real life data, we never have access to the full set of variables involved in explaining the reality we are studying. For instance, in psychology, there are no currently existing data sets that include all of the variables that are known to influence human beings behaviors. And there is also a potentially long list of variables that do influence it without us knowing about it yet.

So, my point was: if the CS systematically reacts (by being significant) to the exclusion of important variables, then it will always be significant in psychological research… And this would invalidate the CS. Any data set that will exclude genes will result in a significant CS, as will any data set that excludes personality, and parenting, and school environment, and work environment, etc.

The problem is that we do not know the real population model with real life data. So, my alternative proposal was that if we picked up a subsample of variables from a simulated data set, this would be highly similar to what we do with real life data. This is a central assumption in this post. If you disagree with it, you will disagree with the full posting.

With real life data, our objective is to come up with the closest possible approximation of the reality on the basis of the variables available in our data set.

So, a good indicator of model fit (be it CS or OFIs) should say that this "best approximation" does provide a good fit to the data.

So, let's suppose we have a data set that only includes X1, X2, X4, Y and Q. And let's also suppose that the previous population model represent the WORLDLY TRUTH about the question we are investigating (we are probably working on a very simple phenomenon, but at least here we can play god with access to the true population model).

SO, the best approximation of the model reality we can come up with is a partial mediation model in which:

y = q + X1 + X4 + v1;

q = x1 + x2+ x4 + v2;

In other words, X1-X2-X4 predict Q, which in turns predicts Y. The relation between X2 and Y is fully mediated through Q, whereas the relation between X1-X4 and Y is only partially mediated through Q (this is to reflect the fact that in the REAL model X1 predicts R which predicts Y and X4 predicts Z which predicts Y). For convenience, let's forget about the regressions between X1->X2->X4 that will anyhow be picked up by the correlations between these variables treated as exogenous. Bottom line: this model is the best we can cook from what we know about the population model. Indeed, when this model is estimated, the results are:

Chi-Square (df) 83.215 (1), p = 0.0000

CFI 0.998 / TLI 0.985

RMSEA 0.091 / RMSEA 90 Percent C. I 0.075 0.108 / SRMR 0.004

MODEL RESULTS

Q ON				
X1	0.298	0.034	8.754	0.000

[Note that in the full model the indirect effect of X1->R->Q is approximately .302]

X2	1.516	0.029	52.630	0.000
X4	1.488	0.017	89.650	0.000
Y ON				
Q	1.575	0.013	120.333	0.000
X1	0.499	0.044	11.226	0.000

[Note that in the full model the indirect effect of X1->R->Y is approximately .985]

X4	0.473	0.033	14.392	0.000

[Note that in the full model the indirect effect of X4->Z->Y is approximately .988]

Residual Variances				
Q	5.615	0.079	70.711	0.000
Y	12.279	0.174	70.711	0.000

STDYX Standardization

Q ON				
X1	0.054	0.006	8.746	0.000
X2	0.393	0.007	52.889	0.000
X4	0.544	0.006	93.140	0.000
Y ON				
Q	0.832	0.006	131.393	0.000
X1	0.048	0.004	11.197	0.000
X4	0.091	0.006	14.377	0.000
Residual Variances				
Q	0.188	0.003	55.501	0.000
Y	0.115	0.002	53.153	0.000

Here we are in a tight position.
The CS is significant, which suggest some form of misfit. And indeed, the parameters estimates are slightly off. The OFI generally suggest good fit however.
BUT, in the present case, with limited data and access only to these variables, this model IS the best approximation of the reality we can come up with. So, John is right, CS reacts to the absence of important variables. But, doesn't real data in the social/educational/psychological sciences always exclude some important variables.
More important however is whether this "incorrect-yet-best-approximation-model" would have resulted in different substantive interpretations. As in our disciplines we do not really interpret results on the basis of the exact size of the path coefficients on a very precise basis, from the preceding results it seems that the substantive interpretations would remain unchanged.
Now, if we do specify an incorrect model, let's say by assuming full mediation for X1 and X4:
y = q + v1;
q = x1 + x2+ x4 + v2;
We obtain the following results.
Chi-Square (df) 402.533 (3), p = 0.0000
CFI 0.990/ TLI 0.976
RMSEA 0.115 / RMSEA CI 0.106 0.125 / SRMR 0.009
MODEL RESULTS

Q	ON				
	X1	0.298	0.034	8.754	0.000
	X2	1.516	0.029	52.630	0.000
	X4	1.488	0.017	89.650	0.000
Y	ON	Q	1.776	0.007	272.298	0.000
Residual Variances					
	Q	5.615	0.079	70.711	0.000
	Y	12.678	0.179	70.711	0.000

STANDARDIZED STDYX MODEL RESULTS

Q	ON				
	X1	0.054	0.006	8.746	0.000
	X2	0.393	0.007	52.889	0.000
	X4	0.544	0.006	93.140	0.000
Y	ON	Q	0.939	0.001	789.879	0.000
Residual Variances					
	Q	0.188	0.003	55.501	0.000
	Y	0.119	0.002	53.265	0.000

Here the OFI would lead to the selection on an incorrectly specified model. However in practice, the CS difference test (as well as the changes in RMSEA that is more than the usual cut offs suggested by Chen et al for instance) would alert us that this model fits less than the preceding one. In addition, it should be noted that this model was not THAT misspecified and did not results in really different substantive conclusions. This is still a mediation model, that is full rather than partial – and note that in the previous partial mediation model the standardized paths between X1-X4 and Y that were taken out are quite small in magnitude (Y ON X1 Stdyx=0.048; Y ON X4 stdyx= 0.091). Yet, the chi CS made a HUGE jump.

I would have like to also be able to compare both of these models with a fully partially mediated model (misspecified by the inclusion of the X2->Y path). But in path analysis this model is just identified.

y = q + X1 + X2 + X4 + v1;

q = x1 + x2+ x4 + v2;

But let's suppose I take out a more substantial standardized path (lets say the Q ON X4 path). Now, the misfit is clear according to every indices:

Chi-Square (df) 5981.695 (2) p = 0.0000

CFI 0. 844 / TLI 0.455/ RMSEA 0. 547 / SRMR 0. 154

Or the Y ON Q path (removing all mediation). Again, except from the CFI and RMSEA, most indices scream that the model is bad.

Chi-Square (df) 987.909 (1) p = 0.0000

CFI 0. 974 / TLI 0.820/ RMSEA 0. 314 / SRMR 0. 074

Now let's suppose a grossly misspecified model:

X1 = q + y + X2 + v1

X2 = X4 + v2;

Chi-Square (df) 4133.231 (3), p = 0.0000

CFI/TLI

CFI 0.772 / TLI 0.468

RMSEA 0.371 / 90 Percent C.I. 0.362 0.381 / SRMR 0.079

Again, all indices (except for SRMR) tell us that this model is bad.

Another one (here a regression model with 2 outcomes, no mediation involved, but preserving the directionality of the predictions):

Y = X1 + X2 + X4 + V1;

Q = X1 + X2 + v2;

Chi-Square (df) 5898.481 (1) , p 0.0000

CFI 0.846/ TLI -0.075

RMSEA 0.768 / 90 Percent C.I. 0.752 0.784 / SRMR 0.164

Again, all indices tell us that this model is bad.

Yes, I know, this example is not convincing one way or the other. Thus, I guess it really reflects reality wouldn't you say?

But remember my claim: These examples apparently show that CS overreacts (and even reacts to absent variables) whereas when the model is misspecified in a realistic way, OFI are there to tell us so.

Part 3. Another example ?

Let's say we have only access to X1 X2 R Q. The "best" model would be (this is now a fully mediated model):

Q = R + X2 + v1

X2 = X1 + v2

R = X1 + V3

The results from this model are (and the CS and OFI both say that the fit is perfect – EVEN when some important variables lacking… This suggest that CS may be erratic?):

Chi-Square (df) 0.398 (2) p = 0.8197

CFI 1.000/ TLI 1.000

RMSEA 0.000 / RMSEA CI 0.000 0.012 / SRMR 0.003

MODEL RESULTS

Q	ON			
	X2	2.981	0.022	136.658	0.000

[BUT this path is way off… Although it will not change much the substantive conclusions]

	R	0.980	0.029	33.265	0.000
X2	ON			
	X1	1.017	0.010	101.424	0.000
R	ON			
	X1	0.303	0.010	29.874	0.000
Residual Variances				
	X2	0.989	0.014	70.711	0.000
	R	1.011	0.014	70.711	0.000
	Q	9.163	0.130	70.710	0.000

STANDARDIZED STDYX MODEL RESULTS

Q	ON			
	X2	0.773	0.004	198.905	0.000
	R	0.188	0.006	32.815	0.000
X2	ON			
	X1	0.712	0.005	144.461	0.000
R	ON			
	X1	0.286	0.009	31.179	0.000
Residual Variances				
	X2	0.493	0.007	70.216	0.000
	R	0.918	0.005	174.676	0.000
	Q	0.307	0.005	60.482	0.000

Now let's take out the X2->Q path from the "best" model:

Chi-Square (df) 3991.975 (2) , p 0.0000

CFI 0.798

TLI 0.393

RMSEA 0.447 / RMSEA 90 Percent C.I. 0.435 0.458 / SRMR 0.634

AGAIN, both the CS and the OFI pick up the misspecification.

Part 4. Following suggestions ?

Some list members suggested that I constrain some paths to specific "wrong" values and that I progressively increase the wrongness of these values to see whether CS reacts before OFI… I did not do this previously as this confirms my point. If CS reacts to slight deviations, it is not useful as these slight deviations would not in the end change the substantive interpretations of the results. But let's take again a look at the first model:

Q	ON			
X1		0.298	0.034	8.754	0.000
X2		1.516	0.029	52.630	0.000
X4		1.488	0.017	89.650	0.000
Y	ON			
Q		1.575	0.013	120.333	0.000
X1		0.499	0.044	11.226	0.000
X4		0.473	0.033	14.392	0.000
Residual Variances				
Q		5.615	0.079	70.711	0.000
Y		12.279	0.174	70.711	0.000

From the previous example, that the X2->Q path is about 1.5. So, restimating the model with this path fixed at this value result in similar fit (bad according to CS and good according to OFI):

Chi-Square (df) 83.524 (2) p 0.0000

CFI	0.998
TLI	0.993

RMSEA 0.064

90 Percent C.I.	0.053	0.076

SRMR 0.005

<u>But fixing this path at 1:</u>

Chi-Square (df) 399.086 (2) P 0.0000

CFI	0.990
TLI	0.964

RMSEA 0.141 / RMSEA 0.129 0.153 / SRMR 0.080

Interestingly, this model (which substantively is about the same), again fits according to most OFI (except RMSEA) but not CS. Here however, both the CS differences tests and the changes in TLI, RMSEA and SRMR converge in telling us that this model fits less than the preceding one. Thus, again, information other than the CS per se allow us to detect this misspecification.

Now, let's start from the right model and constrain Q and Y disturbances to equality.
Chi-Square (df) 1576.639 (2)p = 0.0000

CFI	0.959
TLI	0.856

RMSEA 0.281 / RMSEA 90 Percent C.I. 0.269 0.292 / SRMR 0.304

This is highly interesting. As in the preceding one, both the CS differences tests and the changes in CFI, TLI, RMSEA and SRMR converge in telling us that this model fits less than the preceding one. Particularly for the CS, the jump is huge and apparently unwarranted from the results (note that this is not the first time that the CS apparently over-reacts) since the disturbances seldom have real substantive interest… Moreover, the other parameter estimates remain approximately the same:

Q	ON				
	X1	0.298	0.043	6.935	0.000
	X2	1.516	0.036	41.692	0.000
	X4	1.488	0.021	71.019	0.000
Y	ON				
	Q	1.575	0.011	140.972	0.000
	X1	0.499	0.038	13.152	0.000
	X4	0.473	0.028	16.861	0.000
Residual Variances					
	Q	8.947	0.089	100.000	0.000
	Y	8.947	0.089	100.000	0.000

Now, I will stop here and let you come up with your own conclusions. John made the data set available at: http://www.hec.unil.ch/jantonakis/demo.xls

So you can all play with it.

I'll come back in some time with simulated latent variables data.

ALEX

Now, realistically, I fulfilled what I announced that that was going to do. I may not be able to answer very diligently people reactions - for lack of time.
Anyhow (a) the data set is public and (b) this post contains all the information you need to verify that there are no mistakes in my models or to estimate competing models; (c) the results are there and speak for themselves, you may disagree with my interpretations but I don't really need to answer this. So I guess now that I provided this I will lay back from the debate and read your posts and attempts to change my mind ☺