APPENDIX: STATA SYNTAX WITH COMMENTS

Harry BG Ganzeboom

Version 1, October 28, 2024

To illustrate the lecture, I have created a dataset FakeData.dta, and implemented some elementary applications of SEM modelling. The fake-data emulate a mediation model, in which all variables are measured with three indicators. The Stata syntax is in red, the comments (!!) in black.

```
use C:\Users\harry\Dropbox\))Teaching\SEM\SEM2023\Data\FakeData.dta"

rename _all , lower

pwcorr zm1 zm2 , obs
```

	zm1	zm2
zm1	1 3818	
zm2	0.3886 3818	1 3818

```
sem (M -> zm1 zm2) , standardized
```

!! model is not identified **

```
sem (M -> zm1@aa) (M -> zm2@aa) , var(M@1) standardized
```

!! the equality constraint @aa make the model identified

!! estimated loadings are B = .6233872   SE .0119603

!! Fit is perfect, L2=0

```
sem (M -> zm1 zm2 zm3) , standardized
```

!! bringing in third indicator makes the model identified

!! factor loadings are: .7298234 .5324733 .3421143

!! very heterogeneous in strength, but all statistically significant

```
estat gof , stat(all)
```

!! fit is still perfect -

```
omegacoef zm1 zm2 zm3
```

!! omega reliability is 0.5547 (for future reference, below)

```
sem (X -> zx1 zx2 zx3) (Y -> zy1 zy2 zy3), standardized
```

!! the factor loadings of M and X are very much the same - this is the way the data wore generated. The latent correlation is estimated

at cov .5557689 .0229296. This is also the strength of the effect X → Y.

```
estat gof , stat(all)
```

L2(8) = 8.9, $p < .35$ NS. Which indicates that the model fits the data well. RMSEA = 0.006. pclose > .05

No surprise, this was the way the FakeData.dts ware generated.

```
sem (X -> zx2 zx3) (Y -> zy2 zy3) (X -> Y), standardized
```

Now we estimate the latent effect X → Y at B=.5907829 SE=.0560012. Note the change in SE: reducing the number of indicator from 3 to 2 and leaving out the best ones does not change the point estimate (much) but increased the uncertainty.

```
sem (X -> zx1 zx2) (Y -> zy1 zy2) (X -> Y), standardized
```

Leaving out the worst indicator does not change the point estimated very much, but produces smaller SE: B = .5616237 SE = .0250518. Notice that keeping in the worst indicators zx3 and zy3 still reduced the SE.

```
sem (x123 -> y123) , standardized
```

x123 and y123 are constructed scales from the three indicators. This observed-variables analysis shows a much reduced correlation: B = .3231408 SE = .0141105.

```
sem (X -> x123@1) (Y -> y123) (X -> Y), standardized
reliability(x123 0.5547) reliability(y123 0.5547)
```

This observed-variables model corrects for attenuation using an assumed level of measurement reliability, which I derived using the Omega method (above). The estimated latent effect is right on target: B = .5825511 SE : .0248636.

```
sem (X -> zx1 zx2 zx3) (Y -> zy1 zy2 zy3) (M -> zm1 zm2 zm3) (X -> M
Y) (M -> Y), standardized
```

This is the mediation model with full measurement. Coefficients in the latent part are: X → M .544 X → Y .2623 M → .545. Notice that from the total effect X → Y about half is mediated.

```
sem (x123 -> y123 m123) (m123 -> y123) , standardized
```

This calculates the same mediation model with observed, constructed variables. Here the total effect is 0.32, of which not even 1/3 is mediated.

```
estat teffect
```

Calculates total and indirect effects.

```
sem (X -> zx2 zx3) (Y -> zy2 zy3) (M -> zm2 zm3) (X -> M Y) (M ->
Y), standardized
```

If we calculate the model with only two (the worst) indicators, the point estimates do not change much, but the SE become wider.

```
sem (X -> zx2@1)  (X ->zx3@bb) (Y -> zy2@1) (Y -> zy3@bb) (M -> zm2@1) (M -> zm3@bb)  (X M -> Y) (X -> M),  var(e.zx2@cc) var(e.zx3@dd) var(e.zy2@cc) var(e.zy3@dd) var(e.zm2@cc) var(e.zm3@dd) standardized
```

Constraining the measurement models to be the same for the three latent variables reduces the SE of the direct effect X → M somewhat.

ANALYSIS OF INCOMPLETE DATA WITH FIML

```
use "C:\Users\harry\Dropbox\))Teaching\SEM\SEM2023\Data\FakeData_with_missings.dta", clear
```

```
rename _all , lower
```

```
pwcorr m1 m2 m3, obs
```

I have created this dataset by randomly removing 20% of all values in all measures. This is the scenario of MCAR (Missing Completely at Random).

```
sem (X -> zx1 zx2 zx3) (Y -> zy1 zy2 zy3), standardized
```

This is the complete cases analysis, N=2101. B = .562 SE = .032

```
sem (X -> zx1 zx2 zx3) (Y -> zy1 zy2 zy3), standardized method(mlmv)
```

This is (all) available cases analysis, N=3818. B = .578 SE = .025. Notice the dramatic decrease of the SE.

MULTI-TRAIT MULTI-METHOD MODELLING

```
use "C:\Users\harry\Dropbox\))Teaching\SEM\SEM2023\Data\issp_2009_sem.dta", clear
```

```
rename _all , lower
```

These data are from the ISSP 2009. Respondents and fathers Occupations are measured with two indicators, ASEI (detailed scale) and OSEI (crude scale). The research question is about random and systematic measurement error in these two scales.

```
sem (F -> zfisei@1) (F -> zfosei@bb)  (R -> zisei@1) (R -> zosei@bb) , standardized
```

```
rename _all, lower
```

```
pwcorr zfisei zfosei zisei zosei , obs
```

	zfisei	zfosei	zisei	zosei
zfisei	1			
	13909			

zfosei	0.7498	1		
	8062	9610		
zisei	0.3133	0.3188	1	
	11425	7609	14119	
zosei	0.2959	0.3322	0.7473	1
	11229	7656	13530	13998

The observed correlations between Father's and Respondent's hoovers around 0.31.

```
sem (F -> zfisei@1) (F -> zfosei@bb)  (R -> zisei@1) (R -> zosei@bb)
, standardized
```

The latent correlation is estimated at .447284, SE= .0125569. Listwise N=6121. Model does not fit: L2(2)=25.9, p < .001.

```
sem (F -> zfisei@1) (F -> zfosei@bb)  (R -> zisei@1) (R -> zosei@bb)
(F -> R), standardized method(mlmv)
```

Estimated on all available data (N=16926), the latent correlation is estimated at 0.411 SE: .009. Model does not fit: L2=37.9.

```
sem (F -> zfisei@1) (F -> zfosei@bb)  (R -> zisei@1) (R -> zosei@bb)
(F -> R), standardized method(mlmv) covar(e.zfisei*e.zisei)
covar(e.zfosei*e.zosei)
```

This is the way to include / correct systematic error or method effects: extra correlation between the same measures of different occupation. The model is not identified.

```
sem (F -> zfisei@1) (F -> zfosei@bb)  (R -> zisei@1) (R -> zosei@bb)
(zeddur) (zlnpinc), standardized method(mlmv)
covar(e.zfisei*e.zisei) covar(e.zfosei*e.zosei)
```

The model becomes identified by including two auxiliary variables: (zeddur) (zlnpinc), which are education and income. The latent correlation is now estimated at 0.4021 SE: .0092. Model does not fit: L2(4)=26.9, but makes me happy. Factor loadings for osei and isei are almost equal (0.86), but the systematic error for isei is only 0.034 (ns), while for osei it is 0.088 (t=4.3). This result suggests that respondents make more systematic errors when answering a showcard (osei) than when answering an open question. Random error is almost the same between the two methods. At the same time, the model illustrates that correcting random measurement error is far more important than taking into account systematic measurement error.