NOTES TO SEM LECTURES (yellow: not explicitly discussed)

Harry BG Ganzeboom

Melbourne, October 29 2017

Lecture 4: Panel models

Panel models refer to data collected at different points in time (waves or sweeps) on the same (or partly the same units). If a panel follows unites that entered the survey at the same age, it is often called a cohort study. Panel data are generally regarded to have important advantages to obtain better information about causal processes than other observational (non-experimental) designs. This is so, because the panel design repairs the two important pre-conditions to derive causal inferences from correlational observations:

- Panel data are unambiguous with respect to causal order (what happens first and what happens later);
- Panel data allow you to control confounders without measuring them: this is achieved by either controlling a lagged variable or by fixed-effects models; either way, you control all stable confounders, even the ones you have NOT measured.

Note that these two feature closely resemble the basic features of the controlled randomized experiment (in medical science: randomized trials), the celebrated tool of causal analysis.

Panel designs also have important DISadvantages:

- It takes a long time before you have your data ("we follow then until they or we die" (not) by Parnes, who started the US Survey on Income and Program Participation (SIPP), but died before he ever analyzed the data).
- Often the panel design (and raising the necessary budgets) causes more pressure upon the researchers than the research questions they want to answer.
- Many panel data suffer from attrition which may or may be selective.
    - To their advantage, panel data measure unit characteristics at earlier waves, which gives much opportunity to model the data MAR.
    - Panel attrition is one good reason to turn to SEM for modelling and use its MLMV facilities.

- Panels are useful to study life course changes, but not social change (Firebaugh[1]). For the latter you are advised to collect replicated samples (asking different people), rather than following the same people. Compromise: refreshing panel in which you replace dropped respondents by similar new entries.

- In a surprising number of uses of panel data, the panel features of panel data are NOT used – models could just as well have been estimated with repeated cross-sections.

- A competitor to panel designs is the retrospective design, in which you obtain the information in earlier time point at a single point of data collection.

    - This is routinely used in stratification research, in particular on earlier (parental, first) occupations.

    - Retrospective designs are not principally different from panel designs, but they take less time to collect the information, and suffer less from attrition.

    - SEM offers an interesting opportunity to model retrospective bias in the multiple indicator case: you assume that the causal process at the latent level is from earlier to current, but at the indicator level from current to earlier (Heckman).

    - Ironically, many panel data sets (including my own) collect retrospective data; panel designs often cause you to rethink your design.

- Panel data can be obtained in 2-, 3- or multiple wave designs. There are important differences between 2-wave and 3-wave panels (see below), but I have yet to be convinced of advantages of 4 waves or more. I believe multiple wave designs are often erroneously motivated by the aim to monitor social change – for which panel designs are less useful than repeated cross-sections.

- Principal advantages of 3-wave panels over 2-wave panels:

    - 3-wave panels allow you to estimate simplex reliabilities (== generalized test-retest reliability).

    - 3-wave panels allow you to test stationarity (== are we looking at a dynamic process that does not change of time?).

    - 3-wave panels allow you to compare, even combine, cross-lagged causation models with (simultaneous) reciprocal causation.

[1] Firebaugh, G. (2008). Seven Rules for Social Research. The Seven Rules are: (1) There should be the possibility of surprise in social research. (2) Look for differences that make a difference, and report them. (3) Build reality checks into your research. (4) Replicate where possible. (5) Compare like with like. (6) Use panel data to study individual change and repeated cross-section data to study social change. (7) Let method be the servant, not the master.

Panels and missing values; split form designs

Being able to handle missing data is a crucial element of panel analysis, because of panel attrition. Note, however, that the ability to handle missing data by MLMV opens up important an alternative to a full panel design: by using a split-form design, you can obtain 3-wave panel data, while still obtaining information from eacht unit only on two occasions. All you need is correlations between T1 and T2, T2 and T3, and T1 and T3. These can be obtained from randomly chosen subsamples without overlap.

Simplex reliability

- Reliability assessment is often associated with the use of multiple indicator measurement and then expressed in Cronbach's alpha:
    - Assumes that all indicators measure a single latent concept with equal random error (so a single mean correlation can represent their inter-correlations).
    - Often referred to as "internal consistency", which actually refers more to validity (all indicators represent a single underlying concept in the same way) than reliability. The procedure rather assumes than assesses internal consistency.
    - Coefficient alpha is actually an estimate of the test-retest correlation. It is downwardly biased of there are very few indicators (such as 2-3) or the homogeneity of correlations is not met.
- A more principled definition or reliability is that of a test-retest correlation. Assuming perfect stability at the latent level, and identical measurement processes at two periods, we can estimate the measurement coefficient as √r.
- Simplex models generalize the test-retest design to the 3-waves design, by assuming that there is no direct effect X1 → X3, but all of it is mediated via X2. This is a very plausible assumption in panel designs. We still have to assume that the measurement process is identical between the three waves, but if so:
    - There is no need to use the multiple indicator measurement design. So the procedure also works for single indicator measurement, as well as for index construction that average multiple indicators.
    - We can test or relax the stationarity at the latent level.

    More generally simplex models are called (hidden) markov models [HMM] – this term often refers to models with discrete latent variables [latent classes].

    Alwin (2007)[2] has estimated simplex reliabilities on a host of (some 200) individual level characteristics on any 3-wave panels that he was able to find, mostly in American National Election Studies.

[2] Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. John Wiley & Sons.

- According to the simplex model, the correlation X1,X3 needs to be smaller than the correlation X1,X2 and X2,X3. If r(X13) = r(X12)*r(X23), it implies that the latent transitions are perfectly stable (1.0). If r(X13) > r(X12)*r(X23), you are in trouble. If r(X13) < r(X12)*r(X23), this is evidence of some measurement error, that can be estimated using the simplex model.

Cross-lagged causation design

The cross-lagged [causation] panel design has been the traditional approach to answer the all-important chicken-egg problem: what causes what?

- The cross-lagged panel model can be estimated by OLS – it is in fact nothing else but two multiple regressions. However, using SEM allows you to examine whether the wave 2 correlations are adequately modelled, which is often not the case. SEM then allows you to include a residual correlation between X2 and Y2, which may change the other model estimates.

- Using SEM for cross-lagged panel models has several advantages:
    - Constrained estimation (=testing equality) of the cross-lagged effects
    - MLMV estimation of data with missing values
    - Testing stationarity when there are more than 2 waves.

- In fact, when including the residual correlation the cross-lagged panel model now is a SUR (Seemingly Unrelated Regression) model, which is supposed to deliver more efficient estimates (smaller SE's) than independent OLS. Erik: but this is only true if the set of predictors is different for the two equations.

- However, while this model may result in a better fit of the observed correlation matrix, there is little substantive motivation for introducing the residual correlation. It seems that there is something influencing X2 and Y2 (controlling for X1 and Y1), but we do not know what it is. Or at best, we know it is not the lagged causation process.

Simulation of the reliability estimates in 3-wave panel data

`See SPSS syntax in the appendix`

- I generate a simplex structure (at the 'latent level') with X1 → X2 = 0.45 and X2 → X3 = 0.60. The correlation X1,X3 = 0.27 ≈ 0.45 * 0.60.

- Then I generate 'measurement effects' for x11 .. x33 that are correlated with their latent variables X1 X2 X3 around 0.70, and are correlated with one another (within concepts) around 0.49. A standard SPSS reliability analysis reveals Cronbach's alpha to be 0.75.

- Then I construct three 'index' variables zxx1 zxx2 zxx3 that represent X1 X2 X3 at measured data. Measured index and latent variable are correlated 0.86.

- In SEM I can recover the simulated numbers using

```
sem (X1 -> X2) (X2 -> X3) (X1 -> zxx1@c) (X2 -> zxx2@c) (X3 ->
zxx3@c), var(e.zxx1@a) var(e.zxx2@a) var(e.zxx3@a) var(X1@1)
```

The model has 0 degrees of freedom. This is it.

- An alternative way to recover the structural parameters would be to estimate the model with all nine 'measured' indicators:

```
sem (X1 -> X2) (X1 X2 -> X3) (X1->x11 x21 x31) (X2->x12 x22 x32)
(X3-> x13 x23 x33),  var(X1@1) standardized
```

There are two interesting features to this model:

- The structural effects are estimated with much more efficiency (smaller SE).
- The simplex assumption (X1 → X3 // X2 = 0) is indeed testable (and holds).

- A third alternative would be to correct the model for attenuation, using the estimates we have obtained from the reliability estimates.

```
sem (X1 -> X2) (X2 -> X3) (X1->zxx1@c) (X2->zxx2@c) (X3-> zxx3@c),
reliability(zxx1 0.75 zxx2 0.75 zxx3 0.75) var(X1@1)
```

We obtain the same structural coefficients, with t-values that are closer to the model with 9 input variables (all indicators) than to the simplex model with 3 index variables. HG: I am puzzled why this model has such small SE and obviously does not penalize you for introducing the reliabilities.

Reciprocal causation design

Reciprocal causation estimates causal coefficients between variables that are co-occurring at the same time.

- This model cannot be estimated with OLS, but SEM can do it.
- Reciprocal causation in panel data is just one instance. Reciprocal causation may also be estimated in cross-sectional data. The feature that makes the model identified is the structure of an instrumental variable. Reciprocal causation between X2 and Y2 is identified, because you assume that you have measured a variable X1 that causes Y2 exclusively via X2 and another variable Y1 that influences X2 exclusively via Y2.
- Observe that in the reciprocal causation model in panel data we often have strong instruments (being the lagged values of the variables of interest) and avoid the situation of 'weak instrumentation'. Still, the reciprocal effects are strongly correlated, and I have come across to many situations in which the H0 that the two effects are equal cannot be rejected.
- Identification of the model become stronger when we add more waves and assume stationarity, as well as when we have more data (so think about the importance of MLMV).

Instantaneous versus lagged causation

At first instance the conceptual difference between the cross-lagged and reciprocal causation models seems to be an assumption about when the causation happens: is it indeed plausible that the situation at wave 1 causes the situation at wave 2 with some time lag, or could this happen within a much smaller time frame than the lag in your design? A substantive motivation for the reciprocal model is that the causal process is a continuous process, on which you take snapshots. In this conceptualization the reciprocal causation can be regarded as an accumulation of everything happened between wave 1 and wave 2 (I learned about this interpretation from Peter Schmidt).

The student Workload-Stress data

I study the relationship between (subjective) workload and stress among students of an international secondary school university-preparation program. The data are collected in a 3-wave panel survey, in which the waves are somewhat unequally spaced: beginning of YR1, end of YR1, end of YR2. The research question is about the causal relationships between Subjective Workload (these are complaints from the students about the program workload (three indicators) and a host of indicators that measure their psychological stress level: what causes what?

- There is of course panel attrition. Note that all 2430 students were approached at all waves. So we have incomplete data with various patterns: 111 011 110 101 100 010 001. Only a minority of 1681 ever participating students took part in all tree waves (111: N=470).
- Reliabilities for Subjective Workload (three items) are around 0.70), for Stress (27 items) around 0.95.
- Simplex estimates of measurement relationships:

```
sem (X1 -> X2) (X2 -> X3) (X1->zwl1@c) (X2->zwl2@c) (X3-> zwl3@c),
var(e.zwl1@a) var(e.zwl2@a) var(e.zwl3@a) var(X1@1)
```

➔ measurement: 0.83

```
sem (Y1 -> Y2) (Y2 -> Y3) (Y1->zstress1@c) (Y2->zstress2@c) (Y3->
zstress3@c), var(e.zstress1@a) var(e.zstress2@a) var(e.zstress3@a)
var(Y1@1)
```

➔ measurement: 0.97

The cross-lagged model:

```
sem (zwl1 zstress1 -> zstress2) (zstress1 zwl1 -> zwl2),
covar(e.zwl2*e.zstress2) method(mlmv)
```

three waves:

```
sem (zwl1 zstress1 -> zstress2) (zstress1 zwl1 -> zwl2) (zwl2
zstress2 -> zstress3) (zstress2 zwl2 -> zwl3),
covar(e.zwl3*e.zstress3) covar(e.zwl2*e.zstress2) method(mlmv)
```

with stationarity constraints:

```
sem (zstress1 -> zstress2) (zstress2 -> zstress3) (zwl1 -> zwl2)
(zwl2 -> zwl3) (zwl1 -> zstress2@b) (zwl2 -> zstress3@b)  (zstress1
-> zwl2@a) (zstress2 -> zwl3@a)  , covar(e.zwl3*e.zstress3)
covar(e.zwl2*e.zstress2) method(mlmv)
```

with some more stationarity constraints:

```
sem (zstress1 -> zstress2@c) (zstress2 -> zstress3@c) (zwl1 ->
zwl2@d) (zwl2 -> zwl3@d) (zwl1 -> zstress2@b) (zwl2 -> zstress3@b)
(zstress1 -> zwl2@a) (zstress2 -> zwl3@a) , covar(e.zwl3*e.zstress3)
covar(e.zwl2*e.zstress2) method(mlmv)
```

with latent variables identical to observed variables:

```
sem (ZSTRESS1->zstress1@1) (ZSTRESS2->zstress2@1) (ZSTRESS3-
>zstress3@1) (ZWL1 -> zwl1@1) (ZWL2 -> zwl2@1) (ZWL3 -> zwl3@1)
(ZSTRESS1 -> ZSTRESS2@c) (ZSTRESS2 -> ZSTRESS3@c) (ZWL1 -> ZWL2)
(ZWL2 -> ZWL3) (ZWL1 -> ZSTRESS2@b) (ZWL2 -> ZSTRESS3@b)  (ZSTRESS1
-> ZWL2@a) (ZSTRESS2 -> ZWL3@a) , covar(e.ZWL3*e.ZSTRESS3)
covar(e.ZWL2*e.ZSTRESS2) method(mlmv) iterate(100) var(e.zstress1@0)
var(e.zstress2@0) var(e.zstress3@0) var(e.zwl1@0) var(e.zwl2@0)
var(e.zwl3@0)
```

with simplex estimate of measurement:

```
sem (ZSTRESS1->zstress1@f) (ZSTRESS2->zstress2@f) (ZSTRESS3-
>zstress3@f) (ZWL1 -> zwl1@g) (ZWL2 -> zwl2@g) (ZWL3 -> zwl3@g)
(ZSTRESS1 -> ZSTRESS2@c) (ZSTRESS2 -> ZSTRESS3@c) (ZWL1 -> ZWL2)
(ZWL2 -> ZWL3) (ZWL1 -> ZSTRESS2@b) (ZWL2 -> ZSTRESS3@b)  (ZSTRESS1
-> ZWL2@a) (ZSTRESS2 -> ZWL3@a)  , covar(e.ZWL3*e.ZSTRESS3)
covar(e.ZWL2*e.ZSTRESS2) method(mlmv) iterate(100) var(e.zstress1@d)
var(e.zstress2@d) var(e.zstress3@d) var(e.zwl1@e) var(e.zwl2@e)
var(e.zwl3@e) var(ZWL1@1) var(ZSTRESS1@1)
```

with correction for attenuation, using known reliabilities:

```
sem (ZSTRESS1->zstress1@f) (ZSTRESS2->zstress2@f) (ZSTRESS3-
>zstress3@f) (ZWL1 -> zwl1@g) (ZWL2 -> zwl2@g) (ZWL3 -> zwl3@g)
(ZSTRESS1 -> ZSTRESS2) (ZSTRESS2 -> ZSTRESS3) (ZWL1 -> ZWL2) (ZWL2 -
> ZWL3) (ZWL1 -> ZSTRESS2@b) (ZWL2 -> ZSTRESS3@b)  (ZSTRESS1 ->
ZWL2@a) (ZSTRESS2 -> ZWL3@a)  , covar(e.ZWL3*e.ZSTRESS3)
covar(e.ZWL2*e.ZSTRESS2) method(mlmv) iterate(100) reliability
(zstress1 0.95 zstress2 0.95 zstress3 0.95) reliability (zwl1 0.69
zwl2 0.69 zwl3 0.69)
```

Now the same steps with reciprocal model.

No measurement error:

```
sem (ZSTRESS1->zstress1@1) (ZSTRESS2->zstress2@1) (ZSTRESS3-
>zstress3@1) (ZWL1 -> zwl1@1) (ZWL2 -> zwl2@1) (ZWL3 -> zwl3@1)
(ZSTRESS1 -> ZSTRESS2) (ZSTRESS2 -> ZSTRESS3) (ZWL1 -> ZWL2) (ZWL2 -
> ZWL3) (ZWL2 -> ZSTRESS2@b) (ZWL3 -> ZSTRESS3@b)  (ZSTRESS2 ->
ZWL2@a) (ZSTRESS3 -> ZWL3@a) , method(mlmv) iterate(100)
```

```
var(e.zstress1@0) var(e.zstress2@0) var(e.zstress3@0) var(e.zwl1@0)
var(e.zwl2@0) var(e.zwl3@0)
```

With simplex reliability estimation:

```
sem (ZSTRESS1->zstress1@c) (ZSTRESS2->zstress2@c) (ZSTRESS3-
>zstress3@c) (ZWL1 -> zwl1@d) (ZWL2 -> zwl2@d) (ZWL3 -> zwl3@d)
(ZSTRESS1 -> ZSTRESS2) (ZSTRESS2 -> ZSTRESS3) (ZWL1 -> ZWL2) (ZWL2 -
> ZWL3) (ZWL2 -> ZSTRESS2@b) (ZWL3 -> ZSTRESS3@b)  (ZSTRESS2 ->
ZWL2@a) (ZSTRESS3 -> ZWL3@a) , method(mlmv) iterate(100)
var(e.zstress1@cc) var(e.zstress2@cc) var(e.zstress3@cc)
var(e.zwl1@dd) var(e.zwl2@dd) var(e.zwl3@dd) var(ZWL1@1)
var(ZSTRESS1@1)
```

(does not converge)

Corrected for attenuation:

```
sem (ZSTRESS1->zstress1@0.97) (ZSTRESS2->zstress2@0.97) (ZSTRESS3-
>zstress3@0.97) (ZWL1 -> zwl1@0.83) (ZWL2 -> zwl2@0.83) (ZWL3 ->
zwl3@0.83) (ZSTRESS1 -> ZSTRESS2) (ZSTRESS2 -> ZSTRESS3) (ZWL1 ->
ZWL2) (ZWL2 -> ZWL3) (ZWL2 -> ZSTRESS2@b) (ZWL3 -> ZSTRESS3@b)
(ZSTRESS2 -> ZWL2@a) (ZSTRESS3 -> ZWL3@a) , method(mlmv) iterate(50)
var(e.zstress1@0.05) var(e.zstress2@0.05) var(e.zstress3@0.05)
var(e.zwl1@0.31) var(e.zwl2@0.31) var(e.zwl3@0.31)
```

With known reliabilities:

```
sem (ZSTRESS1->zstress1@1) (ZSTRESS2->zstress2@1) (ZSTRESS3-
>zstress3@1) (ZWL1 -> zwl1@1) (ZWL2 -> zwl2@1) (ZWL3 -> zwl3@1)
(ZSTRESS1 -> ZSTRESS2) (ZSTRESS2 -> ZSTRESS3) (ZWL1 -> ZWL2) (ZWL2 -
> ZWL3) (ZWL2 -> ZSTRESS2@b) (ZWL3 -> ZSTRESS3@b)  (ZSTRESS2 ->
ZWL2@a) (ZSTRESS3 -> ZWL3@a) , method(mlmv) iterate(50) reliability
(zstress1 0.95 zstress2 0.95 zstress3 0.95) reliability (zwl1 0.69
zwl2 0.69 zwl3 0.69)
```

(However, compare to standardized model)

Adding in stationary cross-lagged effects:

```
sem (ZSTRESS1->zstress1@1) (ZSTRESS2->zstress2@1) (ZSTRESS3-
>zstress3@1) (ZWL1 -> zwl1@1) (ZWL2 -> zwl2@1) (ZWL3 -> zwl3@1)
(ZSTRESS1 -> ZSTRESS2) (ZSTRESS2 -> ZSTRESS3) (ZWL1 -> ZWL2) (ZWL2 -
> ZWL3) (ZWL2 -> ZSTRESS2@b) (ZWL3 -> ZSTRESS3@b)  (ZSTRESS2 ->
ZWL2@a) (ZSTRESS3 -> ZWL3@a) (ZWL1 -> ZSTRESS2@bbb) (ZWL2 ->
ZSTRESS3@bbb)  (ZSTRESS1 -> ZWL2@aaa) (ZSTRESS2 -> ZWL3@aaa) ,
method(mlmv) iterate(50) reliability (zstress1 0.95 zstress2 0.95
zstress3 0.95) reliability (zwl1 0.69 zwl2 0.69 zwl3 0.69)
standardized
```

(good fit, but parameters make no sense)

Summary of problems with the reciprocal effects panel model

- Ultimately, the reciprocal effects model with estimation of (simple) reliabilities does not converge. It does converge with corrections-for-attenuation.
- In the Student-Workload-Stress data, I obtain significant and different estimates of the two effects; however, when tested on equality, the effects are not significantly different.
- Neither the cross-lagged nor the reciprocal-effects causation model is very useful to introduce further exogenous variables. You may learn about the direction of causality, but that is about it.

Appendix: SPSS simulation of how measurement unreliability affect estimates of causal relation

```
GET

FILE='C:\Users\Harry\AppData\Local\Temp\issp_2013_2014_NL_def.sav'.

compute X1 = Normal(1).
compute X2 = X1 + normal(2).
compute X3 = X2 + normal(3).

corr x1 x2 x3.

desc x1 x2 x3 /save.

corr zx1 zx2 zx3.

compute x11 = ZX1 + normal(1).
compute x21 = ZX1 + normal(1).
compute x31 = ZX1 + normal(1).

compute x12 = ZX2 + normal(1).
compute x22 = ZX2 + normal(1).
compute x32 = ZX2 + normal(1).

compute x13 = ZX3 + normal(1).
compute x23 = ZX3 + normal(1).
compute x33 = ZX3 + normal(1).

reli var=x11 x21 x31 /summ=all corr.
reli var=x12 x22 x32 /summ=all corr.
reli var=x12 x22 x32 /summ=all corr.

compute xx1=mean(x11,x21,x31).
compute xx2=mean(x12,x22,x32).
compute xx3=mean(x13,x23,x33).

corr xx1 xx2 xx3.

desc xx1 xx2 xx3 /save.

save translate outfile="U:\)Teaching\SEm\Melbourne\simul333.dta"
  /keep=zxx1 zxx2 zxx3 x11 x21 x31 x12 x22 x32 x13 x23 x33
  /type=stata /replace.

corr x1 xx1 x11 x21 x31.
```