# Why we may not need SEM after all

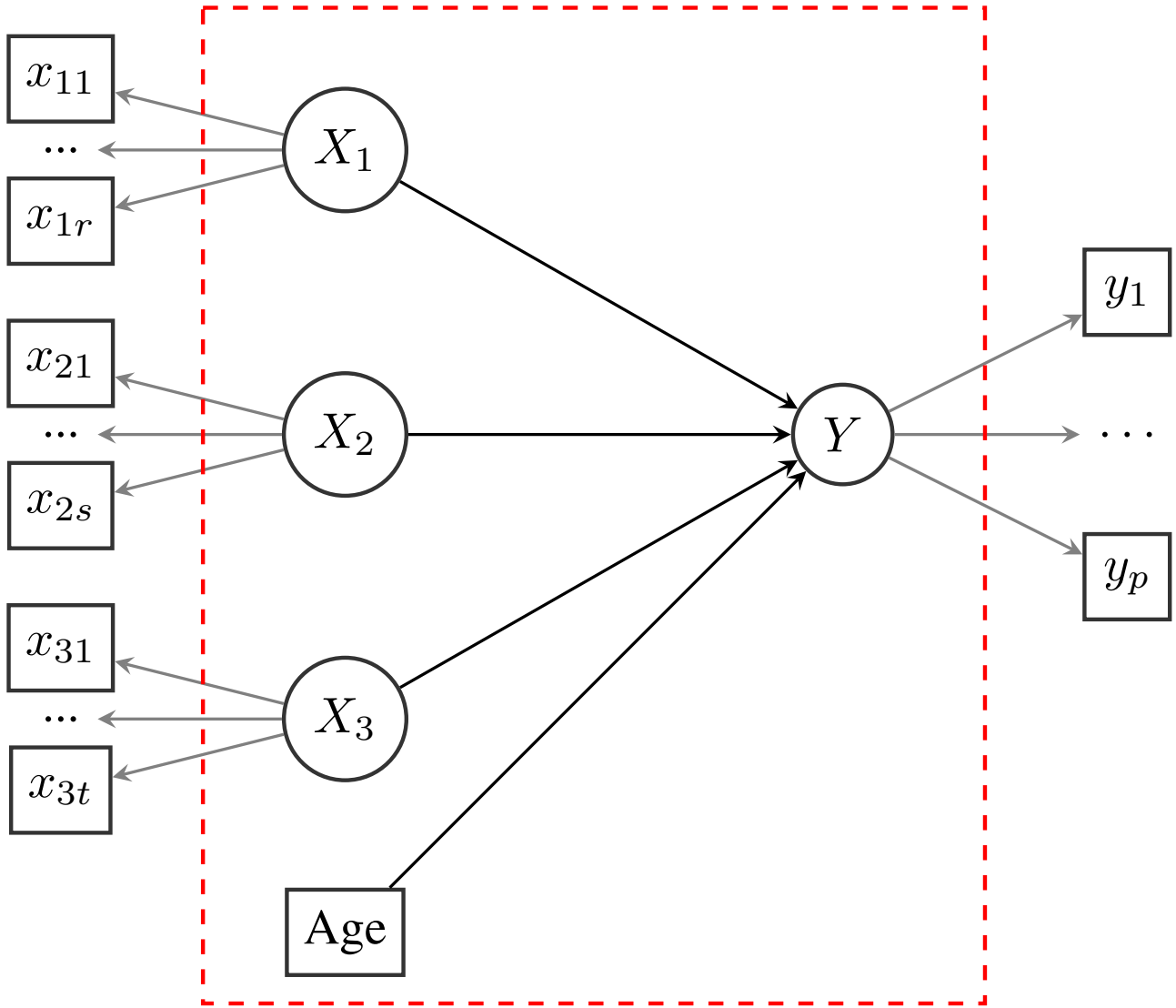
Yves Rosseel & Ines Devlieger
Department of Data Analysis
Ghent University – Belgium

March 15, 2018
Meeting of the SEM Working Group – Amsterdam

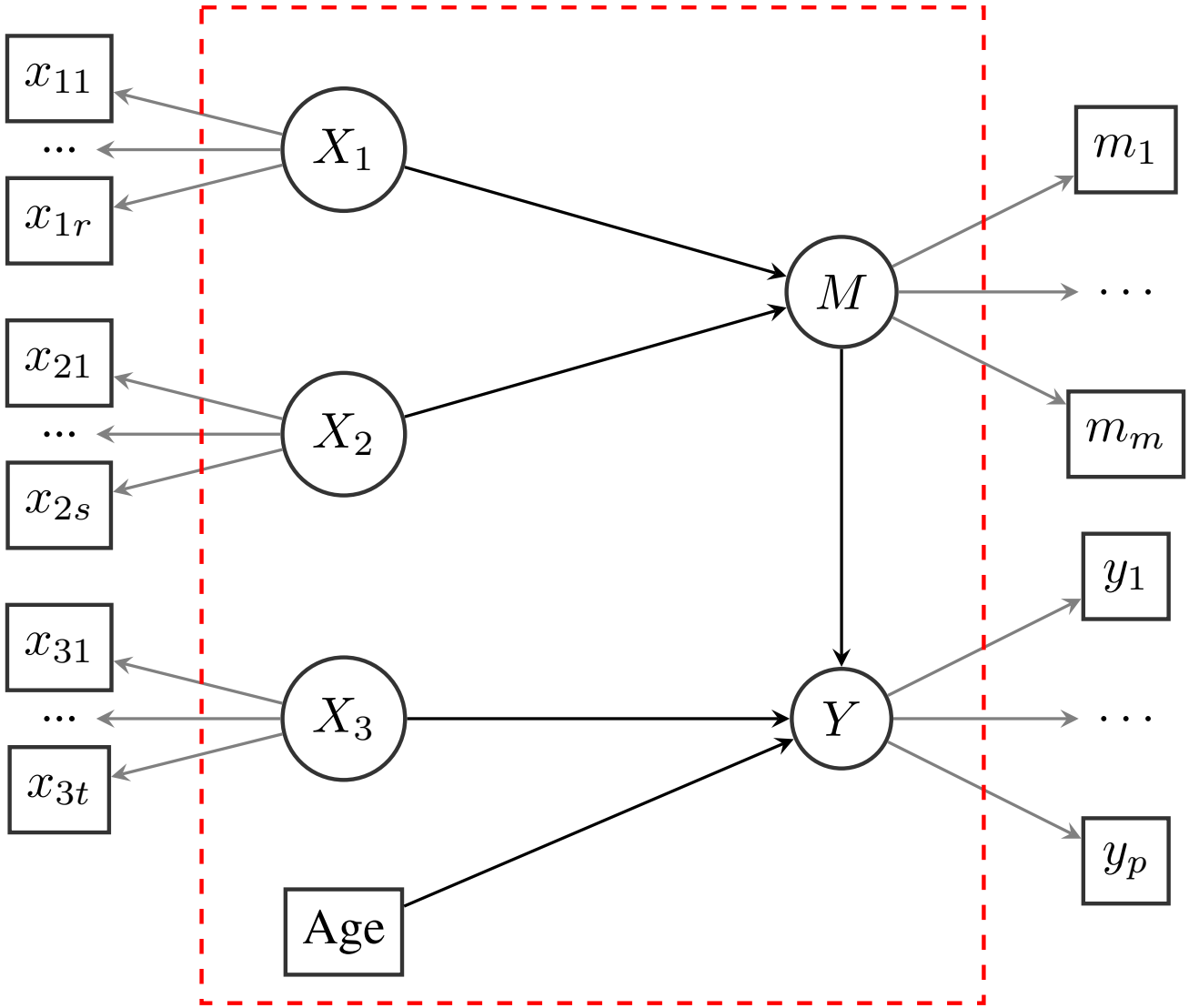
## background

- a typical dataset in the social and behavioural sciences:
  - many constructs (motivation, ability, personality traits, . . . )
  - each construct is measured by a set of (observed) indicators
  - many 'background' variables (age, gender, . . . )
  - (multilevel data, missing data, categorical data)
- the measurement instruments for the latent variables are well established, and usually fit (reasonably) well
- the main focus of the study is the structural part of the model:
  - regression model: variables are either dependent or independent
  - path analysis model: includes mediating effects, perhaps non-recursive
- the sample size is not always very large

# structural model: regression model



# structural model: path analysis model



## the golden standard: structural equation modeling (SEM)

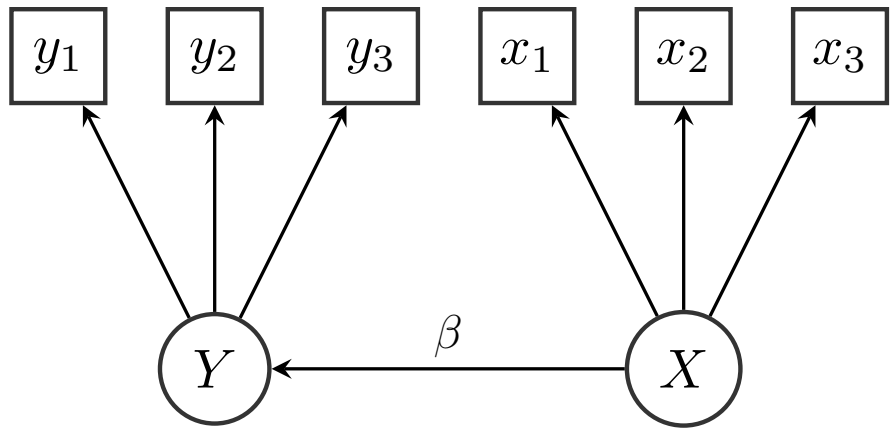
- what do I mean with 'SEM':
  - statistical model: measurement part + structural part
  - estimation procedure: system-wide
    - all the (free) parameters are estimated simultaneously
    - in the continuous case, the software default is usually ML
  - assessment of model fit: global fit measures
- but, what about:
  - model misspecification
  - local fit measures
  - conceptual distinction: measurement part versus structural part
  - small samples
  - ...

## why we may not need SEM after all: alternatives

- alternative approaches:
  - consistent PLS (Dijkstra, T.K., 2010, 2014)
  - model-implied instrumental variables estimation (Bollen, 1996, 2001) (software: R package 'MIIVsem')
  - two-step approaches
  - factor score regression
- shared advantages:
  - reduced model complexity
  - consistent estimates (at least for the structural part)
  - robust to local misspecifications
  - (almost) no convergence issues
  - ...

## a simple example

- consider the regression of a measured latent variable $Y$ on another measured latent variable $X$:




- we are mainly interested in the question: is there a significant effect from $X$ on $Y$? We want to test the hypothesis:

$$H_0 : \beta = 0$$

## data generation

```
> library(lavaan)
> pop.model <- '
+     # factor loadings
+     Y =~ 1*y1 + 0.8*y2 + 0.6*y3
+     X =~ 1*x1 + 0.8*x2 + 0.6*x3
+
+     # regression part
+     Y ~ 0.25*X
+ '
> set.seed(1234)
> Data <- simulateData(pop.model, sample.nobs = 200L, empirical = TRUE)
```

## the golden standard: SEM

```
> model <- '
+     # factor loadings
+     Y =~ y1 + y2 + y3
+     X =~ x1 + x2 + x3
+
+     # regression part
+     Y ~ X
+ '
> fit.sem <- sem(model, data = Data, estimator = "ML")
```

# output SEM

```
> parameterEstimates(fit.sem, add.attributes = TRUE, ci = FALSE)[1:7,]

Parameter Estimates:

  Information                                 Expected
  Information saturated (h1) model          Structured
  Standard Errors                             Standard

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y =~
    y1                1.000
    y2                0.800    0.161    4.972    0.000
    y3                0.600    0.123    4.881    0.000
  X =~
    x1                1.000
    x2                0.800    0.169    4.735    0.000
    x3                0.600    0.129    4.661    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y ~
    X                 0.250    0.114    2.189    0.029
```

## two-step estimation

- old idea:
  - Burt, R.S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological methods & research, 5*, 3–52
  - Anderson, J.C., & Gerbing, D.W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin, 103*, 411–423
- recently, these ideas have been used in the latent class literature, e.g.:

  Bakk, Z., Oberski, D.L., & Vermunt, J.K. (2014). Relating latent class assignments to external variables: standard errors for correct inference. *Political analysis, 22*, 520–540.
- forthcoming: joint work with Zsuzsa Bakk, Jouni Kuha & Yves Rosseel: two-step approach for SEM (with correct inference)

## two-step estimation

- procedure:
  - step 1a: estimate the measurement models for $Y$
  - step 1b: estimate the measurement models for $X$
  - step 2: keeping the parameters of the measurement models fixed to their estimated values, estimate the remaining parameter of the structural part ($\beta$)
  - adjust the standard error(s) of the structural parameters, taking the uncertainty of the first step(s) into account (based on pseudo-ML literature, see Gong & Samaniego, 1981)
- the first steps could be done with any SEM software; for the standard errors, you need custom software
  - a new function called `twostep` has been added to lavaan (0.6)

# two-step estimation in lavaan

```
> fit.twostep <- twostep(model, data = Data)
> parameterEstimates(fit.twostep, add.attributes = TRUE, ci = FALSE)[1:7,]
```

```
Parameter Estimates:

  Information                                 Expected
  Information saturated (h1) model          Structured
  Standard Errors                             External

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y =~
    y1                1.000
    y2                0.800    0.167    4.790    0.000
    y3                0.600    0.126    4.772    0.000
  X =~
    x1                1.000
    x2                0.800    0.176    4.545    0.000
    x3                0.600    0.132    4.545    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y ~
    X                 0.250    0.113    2.208    0.027
```

## factor score regression ('fsr')

- simple idea: replace each latent variable by factor scores
- create a new dataset containing those factor scores
- run a regression analysis (or path analysis) using those factor scores
- widely used in practice
- problems:
  - we treat the factor scores as if they were observed
  - the estimated (structural) parameters will be biased
  - statisticians don't like it, and they will tell applied researchers they should use SEM

## factor score regression (naive version)

- we replace the latent variables by factor scores:

```
> fit.Y <- sem('Y =~ y1 + y2 + y3', data = Data)
> fsY <- lavPredict(fit.Y)

> fit.X <- sem('X =~ x1 + x2 + x3', data = Data)
> fsX <- lavPredict(fit.X)
```

- we fit a simple regression model using these factor scores:

```
> fit.fs <- lm(fsY ~ fsX)
> round(summary(fit.fs)$coefficients[2,], 3)

 Estimate Std. Error    t value   Pr(>|t|)
    0.170      0.073      2.329      0.021
```

- bias:
  - downward bias for the point estimate (about 32%)
  - downward bias for the standard error (about 36%)
- the effect is still significant!

## factor score regression: recent developments

Croon, M. (2002). *Using predicted latent scores in general latent structure models*. In Marcoulides, G., Moustaki, I. (Eds.), Latent variable and latent structure modeling (pp. 195–223). Mahwah, NJ: Lawrence Erlbaum.

Hoshino, T., & Bentler, P.M. (2013). *Bias in factor score regression and a simple solution.* In de Leon, A.R., & Chough, K.C. (Eds.). Analysis of Mixed Data: Methods & Applications. New York: Chapman and Hall/CRC

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement, 76*, 741–770.

Devlieger, I., & Rosseel, Y. (2017). Factor Score Path Analysis. *Methodology, 13*, 31–38.

Takane, Y., & Hwang, H. (2017). Comparisons among several consistent estimators of structural equation models. *Behaviormetrika* (online preprint)

## factor score regression in lavaan

- in lavaan (0.6), factor score regression can be done with the function `fsr()`
- automates the steps required to perform factor score regression (or path analysis) using Croon's correction:

```
> fit.fsr <- fsr(model, data = Data, se = "standard", output = "lavaan")

> parameterEstimates(fit.fsr, add.attributes = TRUE, ci = FALSE)[1,]

Parameter Estimates:

  Information                                 Observed
  Observed information based on                Hessian
  Standard Errors                             Standard

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y ~
    X                 0.250    0.071    3.536    0.000
```

- no bias!
- but standard error is too small

## factor score regression: getting the standard errors right

- an ad-hoc solution was proposed in Devlieger et. al. (2016), but we need a more general solution

  1. the bootstrap
     - works very good
     - intensive, takes time
  2. robust (sandwich type) standard errors
     - the standard approach needs a huge ACOV matrix
  3. correction for a two-step estimation procedure
     - based on the pseudo ML literature (Gong & Samaniego, 1981)
     - not trivial to implement in our framework

- work in progress

## advantages of the ‘fsr’ approach

- consistent point estimates for the structural part of the model
- reduction in model complexity
- the ‘fsr’ approach can handle:
  - missing values for indicators (factor scores are always complete)
  - (in principle) categorical indicators (IRT)
- in contrast to ‘system-wide’ estimators (like maximum likelihood) the ‘fsr’ approach is robust against (local) model misspecifications
- conceptual: strict distinction between measurement model(s) and structural model
- (almost) no convergence issues

## future plans and challenges

- challenge: (analytical) standard errors that perform well in the presence of missing indicators and/or non-normal (but continuous) indicators
- challenge: categorical indicators
- challenge: nonlinear/interaction effects (involving latent variables)
- challenge: models where the distinction between the measurement part and the structural part of the model is not clear
- solved: extension to multilevel SEM (see talk by Ines on EAM in Jena)
- future plans: study the relationship with other related approaches:
  - consistent PLS
  - model-implied instrumental variables estimation
  - two-step approaches
  - ...

# Thank you!