Structural Equation Modeling: models, software and stories

Yves Rosseel
Department of Data Analysis
Ghent University – Belgium

useR!2017 – Brussels
5 July 2017

Yves Rosseel?

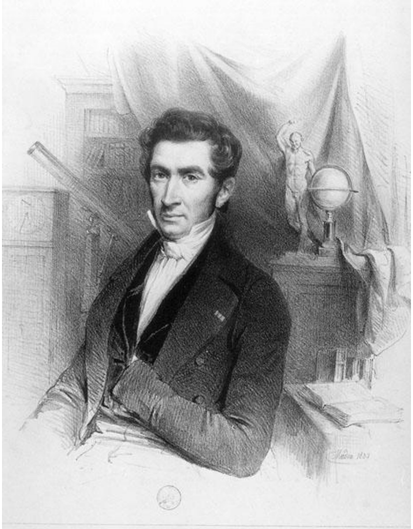
- last year at useR!2016, the opening keynote speaker was Rick Becker ('Forty years of S')
- this year, you have me ('7 years of lavaan')
- who am I anyway?
  - training: experimental psychology
  - phd/postdoc: mathematical psychology
  - job: professor at Ghent University, Faculty of Psychology and Education Sciences, Department of Data Analysis
  - teaching: psychometrics, analysis of repeated measures, applied statistics, . . .
  - (current) research: structural equation modeling
- I wrote an R package ('lavaan')
- I am a Belgian

my social media details:

Ghent University?

- established in 1817 by King William I of the Netherlands
  - 200 years old!
  - in 1930: the first Dutch-speaking university in Belgium
  - state university, 41000 students, 9000 staff members
- rankings?
  - Shanghai ranking (62), Leiden ranking (94), Times Higher Eduction (118), QS World University rankings (125)
  - QS World Rankings by subject: ‘Veterinary Science’ (20), ‘Psychology’: (34), ‘Sports-related’ (41), ‘Agriculture & Forestry’ (46)
- Nobel prizes: Heymans (1938, Medicine); Maeterlinck (1911, Literature)
- famous people? Leo Baekeland (bakelite), Joseph Guislain, Walter Fiers, Marc Van Montagu, Peter Piot, . . . and Adolphe Quetelet
- the best thing about Ghent university? it is in Ghent

Adolphe Quetelet (Gent, 1796–1874)



- PhD at Ghent University in 1819
- astronomer, mathematician, statistician and sociologist
- introduced the 'Body Mass Index' (BMI)
- founded and directed the Brussels Observatory (KMI)
- introduced statistical methods to the social sciences
- "average man" (l'homme moyen): the mean of many 'normally distributed' variables

structure of this talk

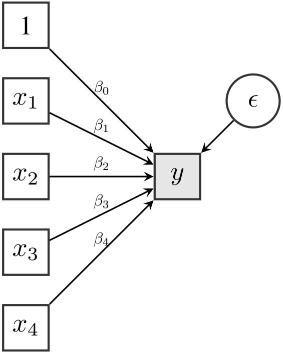
- what is SEM?
- software for SEM (1970 – now)
- the R package ‘lavaan’
  - how did it start? (the origin of the =~ operator)
  - the years after the first public release (2010)
  - lavaan today
  - the lavaan ecosystem
  - growing pains
  - why do we keep doing this?
- last slide

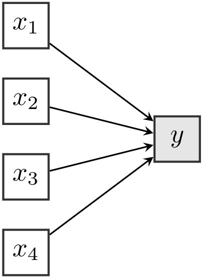
What is SEM?

SEM = structural equation modeling

- SEM is a multivariate statistical modeling technique
- SEM allows us to test a hypothesis/model about the data
  - we postulate a data-generating model
  - this model may or may not fit the data
- what is so special about SEM?
  1. the model may contain latent variables
     - latent variables can be hypothetical 'constructs' (eg., depression) measured by a set of indicators
     - latent variables can be random effects (eg., random intercepts)
     - ...
  2. SEM allows for indirect effects (mediation), reciprocal effects, ...
  3. the model is depicted as a diagram

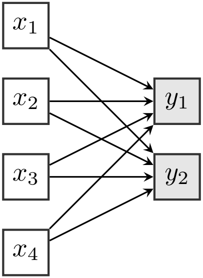
univariate linear regression





$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i \quad (i = 1, 2, \ldots, n)$$

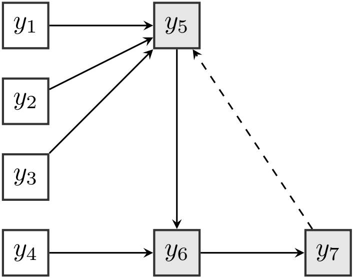
multivariate regression



- strict distinction between ‘dependent’ variables and ‘independent’ variables

SEM: path analysis

- all variables are observed (manifest)
- we allow for indirect effects (eg., of y_5, via y_6 on y_7)
- we allow for cycles (eg. y_7 could influence y_5)



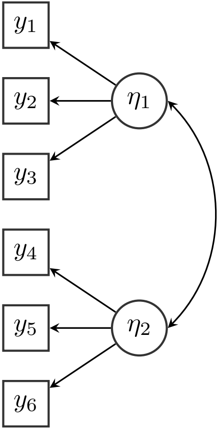
y_5 = reading motivation

y_6 = reading frequency

y_7 = reading ability

SEM: confirmatory factor analysis (CFA)

- measurement model: representing the relationship between one or more latent variables and their (observed) indicators

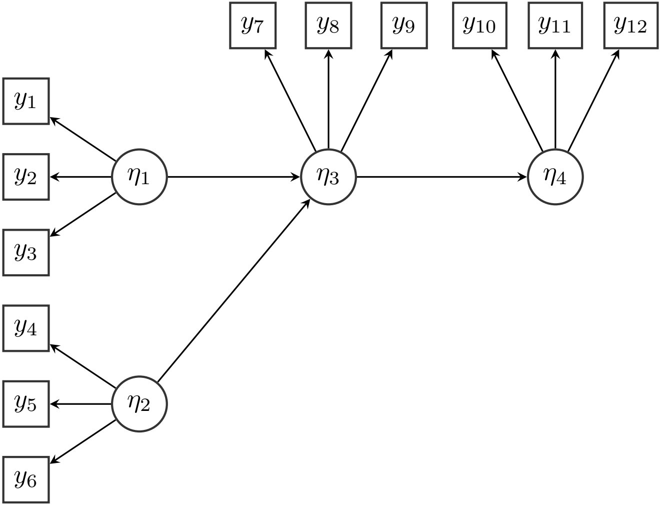


η_1 = depression

η_2 = neuroticism

SEM: measurement models + structural part

- path analysis with latent variables

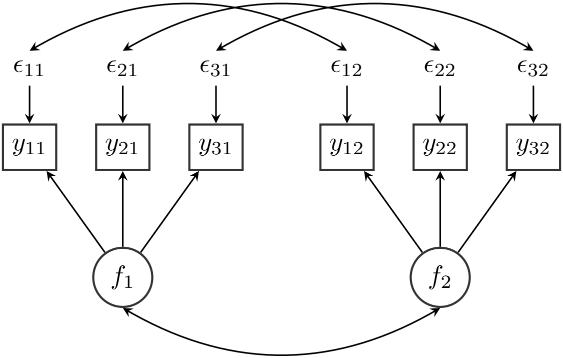


who is using SEM?

- it is widely used in the social sciences
- it is increasingly ‘discovered’ by other fields:
  - medical sciences
  - neuroimaging
  - biology, ecology (climate change!)
  - operation research
  - ...
- SEM software is also used to perform standard analyses (eg., regression), but where there is need for:
  - dealing with missing data
  - robust standard errors, diagnostics
  - (in)equality constraints
  - ...

example: paired t-test using latent variables

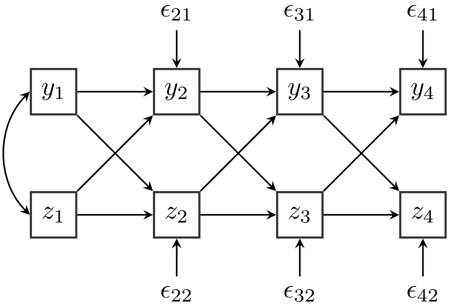
- example with 2 time points:



time 1 time 2

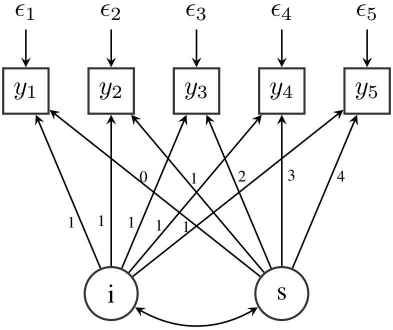
example: panel models with crosslagged effects

- what is the directional effect of one variable on the other?
  - do the two variables develop independently of each other?
  - or does Y exert a greater influence on Z, or vice versa?



example: growth curve model

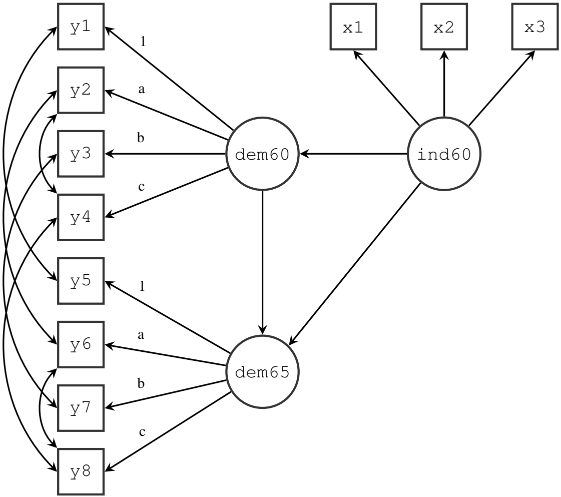
- random intercept and random slope



- y_t = intercept + slope*time + error

example: political democracy

- influence of ‘industrialization (’60)’ on ‘political democracy’ (’60 and ’65)



advantages of SEM

- confirmatory approach: test your theory
- goodness-of-fit measures
- flexible statistical modeling approach
- SEM can handle:
  - missing data (fiml, multiple imputation)
  - (in)equality constraints
  - categorical data (binary, ordinal, count, . . .)
  - discrete and continuous latent variables
  - clustered (multilevel) data
  - . . .
- many other approaches turn out to be special cases (eg., generalized linear mixed models)

disadvantages of SEM

- the modeling flexibility can be overwhelming
- you need dedicated software (not available in, say, SPSS)
- ‘specifying’ your model (using software) can be challenging
- challenges for SEM as a statistical field:
  - better inference for small samples
  - outlier-robust methods are not part of the standard SEM toolbox
  - semiparametric and nonparametric approaches
  - ‘full information’ estimation is computationally often too heavy (we need to integrate out the latent variables)
  - . . .
- we need to better connect with other branches in statistics (graphical models, causal inference, mixed models, . . .)

Software for SEM

software for SEM: commercial – closed-source

- the big four (and the main developer):
  - LISREL ('70s, Karl Jöreskog)
  - EQS ('80s, Peter Bentler)
  - AMOS ('90s, James Arbuckle)
  - Mplus (Bengt Muthén, 1998-now)
- SAS/Stat: proc CALIS, proc TCALIS
- Statistica (SEPATH), Systat (RAMONA), Stata 12
- Mx (Michael Neale, free, closed-source, '90s)
- what about SPSS?
  - SPSS bought AMOS and sells it as a separate product
  - SPSS is bought by IBM (quote from the AMOS website:)

    What it can do for your business

software for SEM: non-commercial – open-source

- outside the R ecosystem:
  - Stata module: ‘gllamm’ (Sophia Rabe-Hesketh, Anders Skrondal, Andrew Pickles, since 2002)
- R packages:
  - sem (John Fox, since 2001)
  - OpenMx (Steven Boker, Michael Neale, … since 2009)
  - lavaan (Yves Rosseel, since 2010)
  - lava (Klaus Holst, since 2012)
- interfaces between R and commercial packages:
  - REQS (Patrick Mair, Eric Wu, since 2008)
  - MplusAutomation (Michael Hallquist, since 2010)

The R package ‘lavaan’

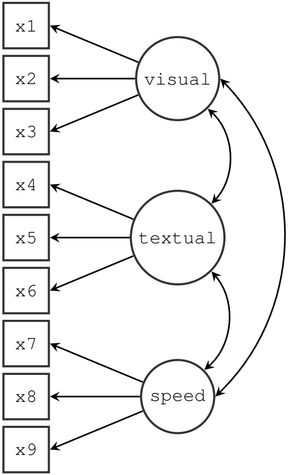
before you start

- many (open-source) statistical software packages (written in R, Julia, Python, . . .) implement *new* statistical ideas
  - you can set your own standards
  - no comparison with other (existing) software
  - no community (yet)
- image writing a package for structural equation modeling (SEM)
  - there is a 'tradition', dating back more than 4 decades
  - there are already many (mostly commercial) software packages available
  - there is already a large community

the beginnings . . .

- the context:
  - in my statistical consultancy years (2000–2008), I often used LISREL, EQS or Mplus, depending on the experience of the client
  - mostly just confirmatory factor analyses (CFA)
  - often very repetitive (same model, multiple datasets)
  - it would be great if we could do everything in R, but (around 2008–2009) the only option was the **sem** package, which was too limited for my purposes
- the initial plan:
  - create a small (private) R package to do only 1 thing: CFA
  - do one thing, do it well (cfr. the Unix philosophy)
  - would be great for teaching too
  - first package (March 2009, never published): **cfa2000**

cfa2000 example: Holzinger & Swineford (1939) 3-factor CFA

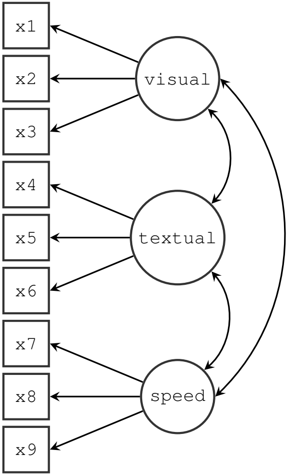


```
library(cfa2000)

# specify 3-factor CFA model
measurement.model <-
    list( visual   = c("x1","x2","x3"),
          textual  = c("x4","x5","x6"),
          speed    = c("x7","x8","x9") )

# fit the model
fit <- cfa(measurement.model = measurement.model,
           data = HolzingerSwineford1939)
summary(fit)
```

cfa2000 partial output

```
Model converged normally after 35 iterations (0.146s)

  Chi-square test full model          85.306
  Degrees of freedom                      24
  P-value                             0.0000

Factor loadings:     Estimate      S.E.   z value  Pr(>|z|)
  visual
     x1                 1.000
     x2                 0.554     0.100     5.554     0.000
     x3                 0.729     0.109     6.685     0.000
  textual
     x4                 1.000
     x5                 1.113     0.065    17.014     0.000
     x6                 0.926     0.055    16.703     0.000
  speed
     x7                 1.000
     x8                 1.180     0.165     7.152     0.000
     x9                 1.082     0.151     7.155     0.000

Factor var/cov:      Estimate      S.E.   z value  Pr(>|z|)
  visual
     visual             0.812     0.146     5.564     0.000
     textual            0.410     0.074     5.552     0.000
     speed              0.263     0.056     4.660     0.000
...
```

cfa2000, August 2009, using formula-like expressions



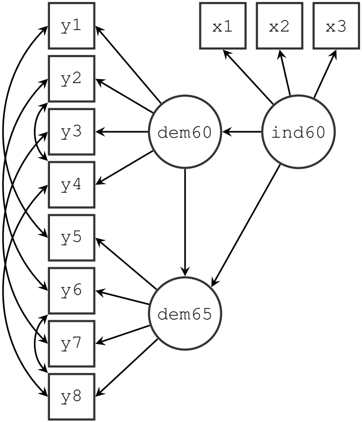
```
visual  =~ x1 + x2 + x3
textual =~ x4 + x5 + x6
speed   =~ x7 + x8 + x9

fit <- cfa(measurement.model = list(visual,
                                    textual,
                                    speed),
           data = HolzingerSwineford1939)

summary(fit)


fit.measures(fit, c("cfi", "rmsea", "srmr"))
```

- nice and easy
- but what about exogenous covariates?
- we need a SEM package after all

second package: ‘semplus’, September 2009

```
Package: semplus
Type: Package
Title: Structural Equation Modeling
Version: 0.9-9
Date: 2009-09-16
Author: Yves Rosseel <yves.rosseel@ugent.be>
Maintainer: Yves Rosseel <yves.rosseel@ugent.be>
Description: Structural Equation Modeling with a formula-based interface
Depends: methods, MASS
License: GPL version 2 or later
LazyLoad: yes
LazyData: yes
Packaged: 2009-10-13 08:18:48 UTC; yves
```

semplus: political democracy example



```
# measurement part
mm <- list(ind60 =~ x1 + x2 + x3,
           dem60 =~ y1 + y2 + y3 + y4,
           dem65 =~ y5 + y6 + y7 + y8)

# correlated errors
ce <- list(y1 ~~ y5,
           y2 ~~ y4 + y6,
           y3 ~~ y7,
           y4 ~~ y8,
           y6 ~~ y8)

# structural part
eqs <- list(dem60 ~ ind60,
            dem65 ~ ind60 + dem60)

fit <- sem(measurement.model = mm,
           eqs = eqs,
           ce = ce,
           data = BollenDemocracy)
```

Jan 2010 – semplus using list() to specify the model

```
model <- list(

   # latent variable definitions
      ind60 =~ x1 + x2 + x3,
      dem60 =~ y1 + y2 + y3 + y4,
      dem65 =~ y5 + y6 + y7 + y8,

   # regressions
      dem60 ~ ind60,
      dem65 ~ ind60 + dem60,

   # residual (co)variances
      y1 ~~ y5,
      y2 ~~ y4 + y6,
      y3 ~~ y7,
      y4 ~~ y8,
      y6 ~~ y8
)
```

March 2010 – semplus using a string literal

```
model <- '
  # latent variable definitions
     ind60 =~ x1 + x2 + x3
     dem60 =~ y1 + y2 + y3 + y4
     dem65 =~ y5 + y6 + y7 + y8

  # regressions
    dem60 ~ ind60
    dem65 ~ ind60 + dem60

  # residual correlations
    y1 ~~ y5
    y2 ~~ y4 + y6
    y3 ~~ y7
    y4 ~~ y8
    y6 ~~ y8
'

fit <- sem(model, data = PoliticalDemocracy)
```

from semplus to lavaan

- the package was named ‘semplus’ because it could do ‘more’ than the sem package
- and it contained the word ‘mplus’
- I contacted the Mplus team (24-02-2010), with some technical questions
- and received an email back (03-03-2010) saying:

> *We own the Mplus trademark. Using the name “semplus” can be construed as a trademark infringement and might also imply our endorsement.*

- eventually, I changed the name to ‘lavaan’ (latent variable analysis)
- lavaan 0.3-1 (about 6470 lines) was released on CRAN on 11 May 2010
  - presented at useR 2010 (NIST, Gaithersburg, Maryland, USA)

the next years

- more and more features were added
- HUGE step: 0.5 added support for categorical data (binary/ordinal)
- more attention for:
  - optimization, scaling, stopping criteria, ...
  - numerical stability, numerical methods
  - what to do if a covariance matrix is not positive-definite?
  - speed
  - ...
- biggest challenge in the early years:
  - the lavaan output was not (always) identical to the output of other (commercial) packages

my program gives (slightly) different results!

- example: Satorra-Bentler scaled test statistic for a 3-factor CFA model using the ‘classic’ Holzinger and Swineford 1939 data (N=301)

program	SB test statistic
lavaan 0.5-22	80.872
Mplus 7.11	81.908
EQS 6.1	81.141
LISREL 8.72	77.396

- experts (often) can not explain these differences
- users of lavaan complained and believed that lavaan’s results could not be trusted

the ‘mimic’ argument

- all fitting functions in lavaan have a `mimic` argument:
  - `mimic="EQS"` to mimic EQS computations
  - `mimic="Mplus"` to mimic Mplus computations
  - `mimic="LISREL"` to mimic LISREL computations (in dev)
  - this was originally intended to convince users that lavaan could produce ‘identical’ results as the (commercial) competition
  - it is now a design goal on its own
- example:

program	SB test stat	lavaan + mimic	SB test stat
lavaan 0.5-22	80.872	`mimic="lavaan"`	80.872
Mplus 7.11	81.908	`mimic="Mplus"`	81.908
EQS 6.1	81.141	`mimic="EQS"`	81.141
LISREL 8.72	77.396	`mimic="LISREL"`	77.396

studying the black box (closed-source) software

- I spent a ridiculous amount of time trying:
  1. to understand (and document) why we observe many subtle (and less subtle) numerical differences in the output of current modern SEM programs
  2. to reproduce results computed by older versions of SEM programs (reproducibility)
  3. to study and compare these (computational and numerical) differences in order to better understand their characteristics
- this is not unlike software archeology
- I learned a lot, and I am still processing the ‘data’
- I discovered the lost art of coding numerical software in an efficient, stable, and elegant way

lavaan today

- current version 0.5-23 (about 43000 lines)
- version 0.6 (including multilevel SEM) is *almost* ready (about 48500 lines)
- the official website:

  http://lavaan.org

- the lavaan paper:

  Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.

- github:

  https://github.com/yrosseel/lavaan

- discussion group (mailing list)

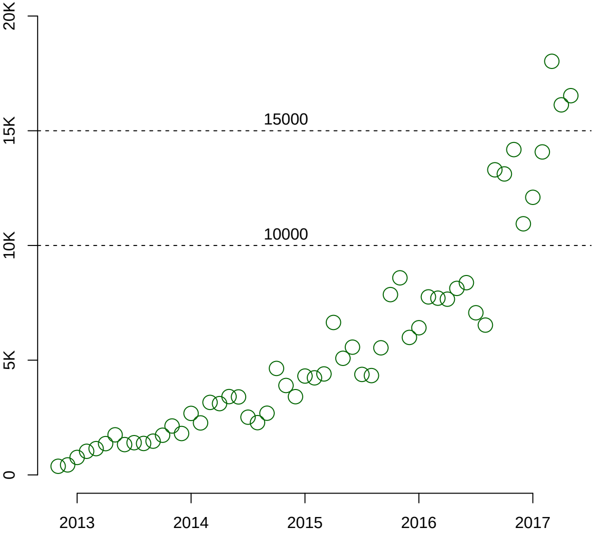
  https://groups.google.com/d/forum/lavaan

how many people use lavaan?

- I have no idea – but it seems like a lot (for a statistical package)
- citations of the lavaan (2012) paper (June 2017):
  - N=1090 (web of science)
  - N=2013 (google scholar)
- lavaan discussion group:
  - N=1502 registered ‘members’ (registration is only needed to post)
  - about 150 posts per month
- cranlogs ‘downloads per months’ suggest an increasing trend:

  https://cranlogs.r-pkg.org/

cranlogs 2012–2017 – per month



the lavaan ecosystem

- **blavaan** (Ed Merkle, Yves Rosseel)

  Bayesian SEM (currently using jags) with a lavaan interface

- **lavaan.survey** (Daniel Oberski)

  survey weights, clustering, strata, and finite sampling corrections in SEM

- **Onyx** (Timo von Oertzen, Andreas M. Brandmaier, Siny Tsang)

  interactive graphical interface for SEM (written in Java)

- **semTools** (Sunthud Pornprasertmanit and many others)

  collection of useful functions for SEM

- **simsem** (Sunthud Pornprasertmanit and many others)

  simulation of SEM models

the lavaan ecosystem (2)

- **semPlot** (Sacha Epskamp)

  visualizations of SEM models

- **EffectLiteR** (Axel Mayer, Lisa Dietzfelbinger)

  using SEM to estimate average and conditional effects

- **nlsem** (Nora Umbach and many others)

  estimation of structural equation models with nonlinear effects and underlying nonnormal distributions

- many others

  bmem, coefficientalpha, eqs2lavaan, fSRM, influence.SEM, MIIVsem, profileR, RAMpath, regsem, RMediation, RSA, rsem, stremo, faoutlier, gimme, lavaan.shiny, matrixpls, MBESS, NlsyLinks, nonnest2, piecewiseSEM, pscore, psytabs, qgraph, sesem, sirt, TAM, userfriendlyscience, . . .

growing pains

- I spend more time ‘testing’ than coding
- each update is somewhat of a nightmare
  - you shall not break a package that depends on your package!
  - you shall not alter the way the output looks!
  - you shall not change the numbers in the output! (same model, same data)
  - you shall not make any mistakes! (lavaan users have come to expect that everything works perfectly, all the time)
  - …
- change (for the better) becomes increasingly difficult
- if only I could start over, with the knowledge I have today

why do we keep doing this?

- why stop:
  - no funding (in Belgium)
  - no support at the faculty/university level: writing software is not in my job description
  - a few (lavaan) users are not very friendly
- why continue?
  - it is (for me) a way to learn about SEM, numerical techniques, statistics, mathematics, . . .
  - it feels more useful than writing yet another paper
  - you meet interesting people
  - open-source (statistical) software is too important
- but at some point (lavaan 1.0?), others will have to take over

last slide: some personal thoughts on R

- lavaan has become a monolithic program, trying to do it all
  - users want consistency
  - essential infrastructure is missing from the core packages
- I find it hard to find ‘gems’ in the R jungle (pieces of code that solve a particular problem in an elegant way)
  - I often feel like I am re-inventing the wheel
  - in a large package, nobody notices if a small part is particularly well done
  - we need a way to ‘mine’ those gems
- we can still learn a lot from other languages (Matlab, Julia, Python, . . .) and/or statistical packages (SAS, Stata, Mplus, . . .)
- I dream of a future language, not unlike Julia, but with a more R-style syntax, called ‘Romeo’

Thank you!

`http://lavaan.org`