Longitudinal Structural Equation Modeling

Yves Rosseel
Department of Data Analysis
Ghent University

Summer School – Using R for personality research
August 23–28, 2014
Bertinoro, Italy

Contents

1 Longitudinal Structural Equation Modeling

1.1 Longitudinal Data Analysis

- longitudinal data analysis is the analysis of *change* in an outcome (or several outcomes) over time
- longitudinal data analysis studies the changes within individuals and the factors that influence change
- longitudinal data is collected in almost every discipline: health, social and behavioural sciences, biological and agricultural sciences, economics, marketing, …
- data is collected in studies using longitudinal designs
- longitudinal data are (by nature) multivariate, and have a complex random-error structure that must be accounted for in the analysis

simple dataset with 4 time points

- 12 subjects, each one has 4 scores on the 'McCarthy Scales of Children's Abilities' (source: Maxwell & Delaney, 2004, Table 11.5)
- 4 time points: 30, 36, 42 and 48 months

```
> MD11.5 <-
+     read.table("http://www.da.ugent.be/datasets/MaxwellDelaney11.5.dat",
+                header=TRUE)
> MD11.5
```

```
   age30 age36 age42 age48
1    108    96   110   122
2    103   117   127   133
3     96   107   106   107
4     84    85    92    99
5    118   125   125   116
6    110   107    96    91
7    129   128   123   128
8     90    84   101   113
9     84   104   100    88
10    96   100   103   105
11   105   114   105   112
12   113   117   132   130
```

data characteristics: the mean structure

- the observed (within-subjects) cell means:

```
> colMeans(MD11.5)

age30 age36 age42 age48
  103   107   110   112
```

- we can consider the 4 time points as the 4 levels of a within-subjects factor 'A'
  - typical research question: is there a (significant) difference between the four time points (on average)?
- or we can consider 'time' (in months) as a continuous variable, with four observed values (30, 36, 42 and 48 months)
  - typical research question: is there a (significant) linear/quadratic/...effect of 'time'?
- and last but not least: what about individual differences? do all subjects follow the same pattern?

data characteristics: the covariance structure

- often, the focus of the analysis is on the means; however, we should also take into account the (complex) correlation structure of the data

```
> round(cor(MD11.5), 2)

      age30 age36 age42 age48
age30  1.00  0.79  0.70  0.60
age36  0.79  1.00  0.76  0.47
age42  0.70  0.76  1.00  0.85
age48  0.60  0.47  0.85  1.00
```

- the variances are important too, so in practice, we need to model the variance-covariance matrix:

```
> round(cov(MD11.5), 1)

      age30 age36 age42 age48
age30 188.0 154.4 127.4 121.2
age36 154.4 200.5 143.6  97.5
age42 127.4 143.6 178.0 168.1
age48 121.2  97.5 168.1 218.0
```

data characteristics: the covariance structure (2)

- typical research questions related to variances/covariances:
  - do the variances change over time?
  - can we detect a particular structure/pattern in the covariance structure?
- modeling the covariances in an adequate way is an important component of longitudinal data analysis

plotting the data

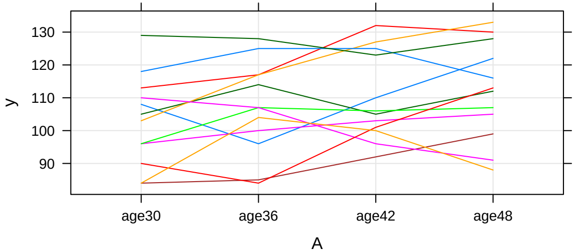
- learn how to plot your data: both average trends, individual trends, ...
- you may need to 'reshape' your data (from wide to long format, or from long to wide format)

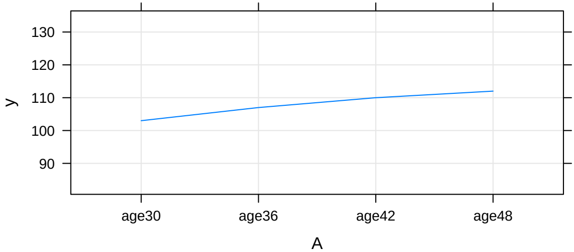
reshape data from ‘wide’ to ‘long’ format + plot

```
> library(reshape)
> MD11.5$subject <- factor(paste("subject",1:nrow(MD11.5),sep=""))
> MD11.5.long <- melt(MD11.5, id.var="subject", variable_name="A")
> MD11.5.long$age <- as.numeric(MD11.5.long$A) - 1
> names(MD11.5.long)[3] <- "y"
> head(MD11.5.long)
```

```
   subject     A   y age
1 subject1 age30 108   0
2 subject2 age30 103   0
3 subject3 age30  96   0
4 subject4 age30  84   0
5 subject5 age30 118   0
6 subject6 age30 110   0
```

```
> library(lattice)
> p1 <- xyplot(y ~ A, groups=subject, data=MD11.5.long, type=c("l","g"))
> p2 <- xyplot(y ~ A, data=MD11.5.long, type=c("a","g"))

> print(p1, split=c(1,1,1,2), more=TRUE); print(p2, split=c(1,2,1,2))
```


130
120
110
100
90
y
age30
age36
age42
age48
A


130
120
110
100
90
y
age30
age36
age42
age48
A

classic analysis: paired t-test

- Student's (1908) paired t-test
- only two time-points (say, age30 and age36)
- we treat 'age' as categorical
- R code:

```
> t.test(MD11.5$age30, MD11.5$age36, paired=TRUE)

        Paired t-test

data:  MD11.5$age30 and MD11.5$age36
t = -1.551, df = 11, p-value = 0.1492
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.676458  1.676458
sample estimates:
mean of the differences
                     -4
```

classic analysis: repeated measures ANOVA

- one of the earliest statistical methods for the analysis of change
- based on the analysis of variance (ANOVA) paradigm, as originally developed by R. A. Fisher
- the mixed-effects ANOVA model (random intercept only):

$$Y_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + b_i + e_{ij}, \quad i = 1, \ldots, N, \ j = 1, \ldots, J$$

  - N is the number of observations; J is the number of repeated measures
  - $\mathbf{X}_{ij}$ is the design matrix; $\boldsymbol{\beta}$ a vector of regression parameters
  - b_i is a random intercept; $b_i \sim N(0, \sigma_b^2)$
  - $e_{ij} \sim N(0, \sigma_e^2)$
- the implied 'compound symmetry' structure for the covariance (of the repeated measures within an observation) is very restrictive

- R code:

```
> fit <- aov(y ~ A + Error(subject), data=MD11.5.long)
> summary(fit)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 11   6624   602.2

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
A          3    552  184.00   3.027 0.0432 *
Residuals 33   2006   60.79
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- corrections (to the degrees of freedom of the F-test) have been proposed by Greenhouse & Geisser (1959) and Huynh & Feldt (1976)
  - $F(1.83, 20.12) = 3.027$, $p = 0.075$ (Greenhouse & Geisser)
  - $F(2.18, 23.94) = 3.027$, $p = 0.063$ (Huynh & Feldt)

sidenote: compound symmetry

- a covariance-matrix has a 'compound symmetry' structure if
    1. all variances are equal: $\mathrm{Var}(Y_l) = \mathrm{Var}(Y_m)$ for every pair (l, m)
    2. all covariances are equal: $\mathrm{Cov}(Y_l, Y_m) = \rho$ for every pair $l \neq m$
- example:

```
> C <- matrix(2, ncol=4, nrow=4)
> diag(C) <- 10
> C

     [,1] [,2] [,3] [,4]
[1,]   10    2    2    2
[2,]    2   10    2    2
[3,]    2    2   10    2
[4,]    2    2    2   10
```

classic analysis: repeated measures MANOVA

- different tradition: repeated measures MANOVA (Box, 1950; Geisser, 1963; Potthof & Roy, 1964)
- based on the multivariate linear model:

$$\mathbf{Y}_i = \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{e}_i$$

- error covariance (of the repeated measures) is 'unstructured' (all elements are freely estimated)
- multivariate tests (Wilks' lambda)
- no random effects
- the original response vector is linearly transformed into a 'difference' vector, reflecting the fact that we are interested in the differences among the responses; different types of transformations are possible

- R code:

```
> d12 <- c(1,-1,0,0)
> d23 <- c(0,1,-1,0)
> d34 <- c(0,0,1,-1)
> M <- cbind(d12,d23,d34)
> M

     d12 d23 d34
[1,]   1   0   0
[2,]  -1   1   0
[3,]   0  -1   1
[4,]   0   0  -1


> fit <- lm( cbind(age30, age36, age42, age48) %*% M ~ 1, data=MD11.5)
> anova(fit, test="Wilks")

Analysis of Variance Table

            Df   Wilks approx F num Df den Df Pr(>F)
(Intercept)  1 0.57251   2.2401      3      9 0.1528
Residuals   11
```

limitations of the classic approaches

- issues with repeated measures ANOVA:
  - correlations among repeated measurement often decay with increasing separation in time
  - the assumption of constant variance across time is often unrealistic
  - complete data only (no missing values)
- issues with repeated measures MANOVA:
  - balanced data only (everyone is measured at the same time points)
  - complete data only (no missing values)
  - no time-varying covariates
- time is treated as an (unordered) categorical variable

the linear mixed model

- the linear mixed model (Laird-ware notation):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \qquad i = 1, 2, \ldots, N$$

- more flexibility in the covariance structure for $\boldsymbol{\epsilon}_i$
- no restrictions on the design matrices ($\mathbf{X}_i$ and $\mathbf{Z}_i$)
- efficient estimation using likelihood-based models (ML and REML)
- can handle unbalanced data, missing data, time-varying covariates, time-invariant covariates
- outside the SEM world, this is the golden standard
- special case: growth curve models (Wishart 1938, Box 1950, Potthof & Roy, 1964)
- can be generalized to account for non-Gaussian responses

- R code:

```
> library(lme4)
> fit.lmer <- lmer(y ~ 1 + age + (1 + age | subject), data=MD11.5.long)
> summary(fit.lmer)

Linear mixed model fit by REML ['lmerMod']
Formula: y ~ 1 + age + (1 + age | subject)
   Data: MD11.5.long

REML criterion at convergence: 345.5595

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 169.85   13.033
          age          14.53    3.812   -0.41
 Residual              34.82    5.901
Number of obs: 48, groups: subject, 12

Fixed effects:
            Estimate Std. Error t value
(Intercept)  103.500      4.023  25.726
age            3.000      1.338   2.241

Correlation of Fixed Effects:
    (Intr)
age -0.475
```

1.2 The SEM approach to longitudinal data analysis

- long history, mostly for 'balanced data': same number of time points for each observation
  - repeated measures models
  - panel models, simplex models, autoregressive models
  - growth curve models (random coefficient models)
  - hybrid models (growth curve + autoregressive)
  - latent-state, latent-trait models
  - latent difference scores models
  - ...
- multilevel SEM
  - combines 'mixed models' with path analysis and latent variables
  - allows for unbalanced data
  - relatively new, active research; major software package: Mplus

1.3 Repeated measures models (using SEM)

- how do the *means* change over time (on average)
- we treat 'time' as a categorical variable with t levels
- SEM version of repeated measures ANOVA
- but much more flexible:
  - the (error) covariance structure is not restricted to compound symmetry
  - we can use latent variables (instead of observed variables), and study the differences between latent means
- if we use latent variables, we first need to establish measurement invariance across time

the naive approach

- what is wrong with this approach?

```
> library(lavaan)
> model <- '
+     age30 ~ i1*1
+     age36 ~ i2*1
+     age42 ~ i3*1
+     age48 ~ i4*1
+
+     age30 ~~ v1*age30
+     age36 ~~ v2*age36
+     age42 ~~ v3*age42
+     age48 ~~ v4*age48
+ '
> fit <- lavaan(model, data = MD11.5)
> lavTestWald(fit, constraints = 'i1 == i2; i2 == i3; i3 == i4')

$stat
[1] 3.058521

$df
[1] 3

$p.value
[1] 0.3826902
```

adding correlated residuals

- we allow the residuals to be correlated:

```
> model <- '
+     age30 ~ i1*1; age36 ~ i2*1; age42 ~ i3*1; age48 ~ i4*1
+
+     age30 ~~ v1*age30
+     age36 ~~ v2*age36
+     age42 ~~ v3*age42
+     age48 ~~ v4*age48
+
+     age30 ~~ age36 + age42 + age48
+     age36 ~~ age42 + age48
+     age42 ~~ age48
+ '
> fit <- lavaan(model, data = MD11.5)
> lavTestWald(fit, constraints = 'i1 == i2; i2 == i3; i3 == i4')

$stat
[1] 8.960391

$df
[1] 3

$p.value
[1] 0.02982217
```

alternative test: LRT

- specify a null model with equal intercepts/means

```
> model.equal <- '
+     age30 ~ i1*1; age36 ~ i1*1; age42 ~ i1*1; age48 ~ i1*1
+
+     age30 ~~ v1*age30
+     age36 ~~ v2*age36
+     age42 ~~ v3*age42
+     age48 ~~ v4*age48
+
+     age30 ~~ age36 + age42 + age48
+     age36 ~~ age42 + age48
+     age42 ~~ age48
+ '
> fit.equal <- lavaan(model.equal, data = MD11.5)
> anova(fit, fit.equal)

Chi Square Difference Test

          Df    AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
fit        0 367.65 374.44 0.0000
fit.equal  3 368.34 373.68 6.6927     6.6927       3    0.08236 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

compound symmetry

- the classical repeated measures ANOVA implies a compound symmetry structure for the residuals:

```
> model <- '
+     age30 ~ i1*1; age36 ~ i2*1; age42 ~ i3*1; age48 ~ i4*1
+
+     age30 ~~ v1*age30
+     age36 ~~ v1*age36
+     age42 ~~ v1*age42
+     age48 ~~ v1*age48
+
+     age30 ~~ c*age36 + c*age42 + c*age48
+     age36 ~~ c*age42 + c*age48
+     age42 ~~ c*age48
+ '
> fit <- lavaan(model, data = MD11.5)
> summary(fit)

lavaan (0.5-17.700) converged normally after  21 iterations

  Number of observations                            12

  Estimator                                         ML
  Minimum Function Test Statistic               17.066
  Degrees of freedom                                 8
```

```
  P-value (Chi-square)                           0.029

Parameter estimates:

  Information                                 Expected
  Standard Errors                             Standard

                   Estimate  Std.err  Z-value  P(>|z|)
Covariances:
  age30 ~~
    age36     (c)   124.069   56.434    2.198    0.028
    age42     (c)   124.069   56.434    2.198    0.028
    age48     (c)   124.069   56.434    2.198    0.028
  age36 ~~
    age42     (c)   124.069   56.434    2.198    0.028
    age48     (c)   124.069   56.434    2.198    0.028
  age42 ~~
    age48     (c)   124.069   56.434    2.198    0.028

Intercepts:
    age30    (i1)   103.000    3.871   26.610    0.000
    age36    (i2)   107.000    3.871   27.643    0.000
    age42    (i3)   110.000    3.871   28.418    0.000
    age48    (i4)   112.000    3.871   28.935    0.000

Variances:
    age30    (v1)   179.792   57.193
    age36    (v1)   179.792   57.193
```

```
    age42     (v1)   179.792   57.193
    age48     (v1)   179.792   57.193


> model.equal <- '
+     age30 ~ i1*1; age36 ~ i1*1; age42 ~ i1*1; age48 ~ i1*1
+
+     age30 ~~ v1*age30
+     age36 ~~ v1*age36
+     age42 ~~ v1*age42
+     age48 ~~ v1*age48
+
+     age30 ~~ c*age36 + c*age42 + c*age48
+     age36 ~~ c*age42 + c*age48
+     age42 ~~ c*age48
+ '
> fit.equal <- lavaan(model.equal, data = MD11.5)
> anova(fit, fit.equal)

Chi Square Difference Test

          Df    AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
fit        8 368.71 371.62 17.066
fit.equal 11 371.47 372.92 25.817      8.751       3    0.03279 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- compare this to:

```
> fit.aov <- aov(y ~ 1 + A + Error(subject), data=MD11.5.long)
> summary(fit.aov)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 11   6624   602.2

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
A          3    552  184.00   3.027 0.0432 *
Residuals 33   2006   60.79
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- here we get an F-statistic, while lavaan is using a chi-square statistic

alternative parameterization, using random intercept

- instead of specifying the compound symmetry structure directly, we can postulate a random effect (random intercept)
- this is the ‘growth curve’ approach

```
> model.A <- '
+     int =~ 1*age30 + 1*age36 + 1*age42 + 1*age48
+
+     # intercepts (fixed effects)
+     int ~ 0
+     age30 ~ i1*1
+     age36 ~ i2*1
+     age42 ~ i3*1
+     age48 ~ i4*1
+
+     # random intercept
+     int ~~ int
+
+     # force same variance for all (compound symmetry)
+     age30 ~~ v1*age30
+     age36 ~~ v1*age36
+     age42 ~~ v1*age42
+     age48 ~~ v1*age48
+ '
```

```
> fit.A <- lavaan(model.A, data=MD11.5)
> summary(fit.A, standardized = TRUE)

lavaan (0.5-17.700) converged normally after  21 iterations

  Number of observations                            12

  Estimator                                         ML
  Minimum Function Test Statistic               17.066
  Degrees of freedom                                 8
  P-value (Chi-square)                           0.029

Parameter estimates:

  Information                                 Expected
  Standard Errors                             Standard

                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Latent variables:
  int =~
    age30             1.000                              11.139    0.831
    age36             1.000                              11.139    0.831
    age42             1.000                              11.139    0.831
    age48             1.000                              11.139    0.831

Intercepts:
    int               0.000                               0.000    0.000
    age30     (i1)  103.000    3.871   26.610    0.000  103.000    7.682
```

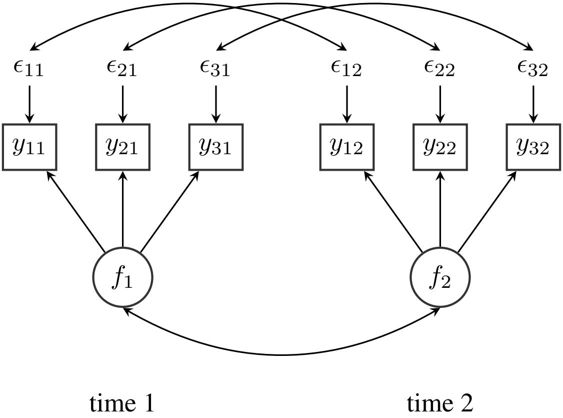
```
    age36    (i2)   107.000    3.871   27.643    0.000  107.000     7.980
    age42    (i3)   110.000    3.871   28.418    0.000  110.000     8.204
    age48    (i4)   112.000    3.871   28.935    0.000  112.000     8.353

Variances:
    int             124.069   56.434                      1.000     1.000
    age30    (v1)    55.722   13.134                     55.722     0.310
    age36    (v1)    55.722   13.134                     55.722     0.310
    age42    (v1)    55.722   13.134                     55.722     0.310
    age48    (v1)    55.722   13.134                     55.722     0.310


> fitted(fit)$cov

      age30   age36   age42   age48
age30 179.792
age36 124.069 179.792
age42 124.069 124.069 179.792
age48 124.069 124.069 124.069 179.792
```

repeated measures ANOVA in SEM

- we can mimic the classical repeated measures ANOVA in a SEM framework
- using two time-points only, this is the SEM equivalent of the paired t-test
- but we can relax the compound symmetry restriction
  - we can allow for an unstructured covariance structure
  - or we could impose an autoregressive AR(1) structure
  - $\ldots$
- but above all, we can replace the observed variables by latent variables

repeated measures using latent variables

- example with 2 time points:



comments

- first of all, we need to establish measurement invariance across time points
  - it is tempting to do this using a multiple group analysis, using the time points as group levels, but this will not allow us to specify correlated residuals among the corresponding variables (and the time points are not independent)
  - therefore, we need to use labels for the different time points (for factor loadings and intercepts of observed variables), and impose the equality constraints by using the same label for the different time points
- since we wish to compare the latent means, we need 'strong invariance':
  - equal factor loadings
  - equal intercepts/means of the observed variables
- usually, we allow the residuals variances of the corresponding variables across time to be correlated

- if we have more than two time points, we can allow for all possible correlations among the repeated latent variables (this corresponds to the ‘unstructured’ assumption)
- the latent mean/intercept of the first time point is fixed to zero, while we estimate the latent mean/intercept of the other time points (although alternative coding schemes are possible)

example

- example from Todd Little's book (Longitudinal SEM, 2013): table 3.8 and figure 3.10 (but with equality constraints)
- the latent variable 'positive affect' is measured by three indicators (Glad, Cheerful and Happy): 823 children in grades 7 en 8 responded to questions like "In the past 2 weeks, I have felt ..." (with 4 response categories: almost never, seldom, often, almost always)
- measured at two time points: in the fall of two successive school years
- main question: is there a significant difference in (self-reported) 'positive affect' between the two time points?
- this is the SEM equivalent of the paired t-test

R code: reading in the sample statistics

```
> MEAN <- c(3.06893, 2.92590, 3.11013, 3.02577, 2.85656, 3.09346)
> SDS  <- c(0.84194, 0.88934, 0.83470, 0.84081, 0.90864, 0.83984)
> lower <- '
+     1.00000
+     0.55226    1.00000
+     0.56256    0.60307    1.00000
+     0.31889    0.35898    0.27757    1.00000
+     0.24363    0.35798    0.31889    0.56014    1.00000
+     0.32217    0.36385    0.32072    0.56164    0.59738    1.00000 '
> COV <- getCov(lower, sds=SDS, names = c("Glad1", "Cheer1", "Happy1",
+                                          "Glad2", "Cheer2", "Happy2"))
> COV

           Glad1    Cheer1    Happy1     Glad2    Cheer2    Happy2
Glad1  0.7088630 0.4135162 0.3953488 0.2257459 0.1863819 0.2278048
Cheer1 0.4135162 0.7909256 0.4476782 0.2684330 0.2892800 0.2717608
Happy1 0.3953488 0.4476782 0.6967241 0.1948053 0.2418595 0.2248294
Glad2  0.2257459 0.2684330 0.1948053 0.7069615 0.4279434 0.3965998
Cheer2 0.1863819 0.2892800 0.2418595 0.4279434 0.8256266 0.4558680
Happy2 0.2278048 0.2717608 0.2248294 0.3965998 0.4558680 0.7053312
```

R code: setting up a longitudinal CFA

```
> model <- '
+     posAffect1 =~ 1*Glad1 + ch*Cheer1 + ha*Happy1
+     posAffect2 =~ 1*Glad2 + ch*Cheer2 + ha*Happy2
+
+     # intercepts
+     Glad1  ~ igl*1
+     Glad2  ~ igl*1
+     Cheer1 ~ ich*1
+     Cheer2 ~ ich*1
+     Happy1 ~ iha*1
+     Happy2 ~ iha*1
+
+     # residual covariances
+     Glad1  ~~ Glad2
+     Cheer1 ~~ Cheer2
+     Happy1 ~~ Happy2
+
+     # latent means
+     posAffect1 ~ 0*1 # baseline
+     posAffect2 ~ 1   # difference compared to baseline
+ '
> fit <- cfa(model, sample.cov = COV,
+            sample.mean = MEAN,
+            sample.nobs = 823)
> summary(fit, fit.measures = TRUE, standardized = TRUE)
```

```
lavaan (0.5-17.700) converged normally after  36 iterations

  Number of observations                           823

  Estimator                                         ML
  Minimum Function Test Statistic               20.279
  Degrees of freedom                                 9
  P-value (Chi-square)                           0.016

Model test baseline model:

  Minimum Function Test Statistic             1761.666
  Degrees of freedom                                15
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.994
  Tucker-Lewis Index (TLI)                       0.989

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -5381.011
  Loglikelihood unrestricted model (H1)      -5370.871

  Number of free parameters                         18
  Akaike (AIC)                               10798.022
  Bayesian (BIC)                             10882.855
```

```
  Sample-size adjusted Bayesian (BIC)        10825.694

Root Mean Square Error of Approximation:

  RMSEA                                          0.039
  90 Percent Confidence Interval          0.016  0.062
  P-value RMSEA <= 0.05                          0.763

Standardized Root Mean Square Residual:

  SRMR                                           0.019

Parameter estimates:

  Information                                 Expected
  Standard Errors                             Standard

                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Latent variables:
  posAffect1 =~
    Glad1             1.000                               0.603    0.715
    Cheer1   (ch)     1.150    0.046   25.063    0.000    0.693    0.780
    Happy1   (ha)     1.076    0.043   25.208    0.000    0.648    0.777
  posAffect2 =~
    Glad2             1.000                               0.607    0.723
    Cheer2   (ch)     1.150    0.046   25.063    0.000    0.698    0.768
    Happy2   (ha)     1.076    0.043   25.208    0.000    0.653    0.780
```

```
Covariances:
  Glad1 ~~
    Glad2              0.032    0.015    2.074    0.038    0.032    0.092
  Cheer1 ~~
    Cheer2             0.017    0.016    1.047    0.295    0.017    0.053
  Happy1 ~~
    Happy2            -0.011    0.014   -0.800    0.424   -0.011   -0.041
  posAffect1 ~~
    posAffect2         0.202    0.021    9.840    0.000    0.553    0.553

Intercepts:
    Glad1    (igl)     3.067    0.027  114.088    0.000    3.067    3.639
    Glad2    (igl)     3.067    0.027  114.088    0.000    3.067    3.652
    Cheer1   (ich)     2.915    0.029   99.814    0.000    2.915    3.283
    Cheer2   (ich)     2.915    0.029   99.814    0.000    2.915    3.204
    Happy1   (iha)     3.123    0.027  115.351    0.000    3.123    3.740
    Happy2   (iha)     3.123    0.027  115.351    0.000    3.123    3.726
    psAffc1            0.000                               0.000    0.000
    psAffc2           -0.040    0.025   -1.617    0.106   -0.066   -0.066

Variances:
    Glad1              0.347    0.022                      0.347    0.489
    Cheer1             0.308    0.024                      0.308    0.391
    Happy1             0.277    0.021                      0.277    0.397
    Glad2              0.336    0.022                      0.336    0.477
    Cheer2             0.340    0.025                      0.340    0.411
    Happy2             0.275    0.021                      0.275    0.392
    posAffect1         0.363    0.029                      1.000    1.000
```

```
posAffect2        0.369    0.030                          1.000    1.000
```

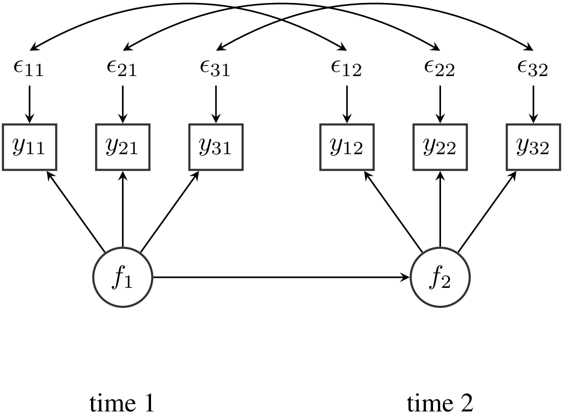
- answer: there is NO significant difference between the two time points

1.4 Panel models for longitudinal data

- panel models postulate *directional* (regression) relationships among the repeated measures
- the 'covariance' is replaced by a 'regression'
- both within repeated variables (autoregressive) and between repeated variables (cross-lagged)
- focus on the model-implied covariance/correlation structure
- the means are usually ignored
- some subtypes:
  - autoregressive models (the simplex model)
  - cross-lagged models
  - latent autoregressive/cross-lagged models
  - ...

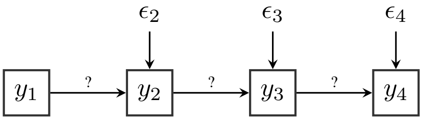
example panel model with a single latent variable

- example with 2 time points:



autoregressive models

- each time point is regressed on a previous time point (first order) , or an even further time point (second order, third order, ...)
- alternative names: Markov models, simplex models, panel models, ...
- earliest development dates back to the seminal work of Guttman (1954)
- example first-order univariate autoregressive model:



The simplex change process

- basic assumption: if individuals are changing at a steady rate, the pattern of correlations from one time point to the next should follow a *simplex pattern*
- key characteristic of a simplex pattern: correlations decrease in magnitude as a function of distance from the diagonal of the correlation matrix
- example of a 'Guttman simplex pattern' (correlation = 0.9):

```
      y1    y2    y3    y4
y1 1.000
y2 0.900 1.000
y3 0.810 0.900 1.000
y4 0.729 0.810 0.900 1.000
```

- example of a 'Guttman simplex pattern' (correlation = 0.5):

```
      y1    y2    y3    y4
y1 1.000
y2 0.500 1.000
y3 0.250 0.500 1.000
y4 0.125 0.205 0.500 1.000
```

- the simplex model assumes that external (eg. environmental or contextual) influences are minimal, and can be ignored
  - typical example in psychology: early cognitive development
  - counterexample: social adjustment in school (often strongly influenced by contextual influences)
- questions we can answer with panel models:
  - are the effects (regression coefficients) the same across time points?
    - makes more sense if the time span between variables is the same
    - if yes, this would suggest that the process does not change over time
  - are the residual variances the same across time points?
    - often unrealistic
  - do we have correlated residuals? if so, are they the same across time points?

example: McCarthy Scales of Children's Abilities

- a first-order autoregressive model

```
> library(lavaan)
> model.age <- '
+     # first order
+     age36 ~ age30
+     age42 ~ age36
+     age48 ~ age42
+ '
> fit <- sem(model.age, data = MD11.5, fixed.x = FALSE)
> summary(fit, standardized = TRUE)

lavaan (0.5-17.700) converged normally after  14 iterations

  Number of observations                            12

  Estimator                                         ML
  Minimum Function Test Statistic                7.661
  Degrees of freedom                                 3
  P-value (Chi-square)                           0.054

Parameter estimates:

  Information                                 Expected
  Standard Errors                             Standard
```

```
                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Regressions:
  age36 ~
    age30             0.821    0.181    4.540    0.000    0.821    0.795
  age42 ~
    age36             0.716    0.177    4.054    0.000    0.716    0.760
  age48 ~
    age42             0.944    0.167    5.669    0.000    0.944    0.853

Variances:
    age36            67.650   27.618                     67.650    0.368
    age42            68.863   28.113                     68.863    0.422
    age48            54.328   22.179                     54.328    0.272
    age30           172.333   70.355                    172.333    1.000
```

```
> round(cor(MD11.5[,c("age30","age36","age42","age48")]), 3)
```

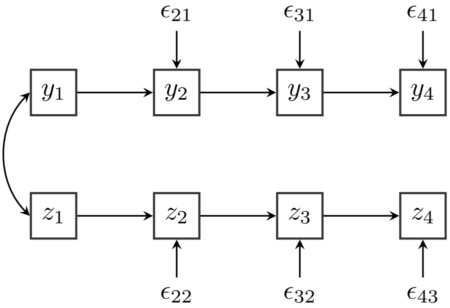
```
      age30 age36 age42 age48
age30 1.000 0.795 0.696 0.599
age36 0.795 1.000 0.760 0.466
age42 0.696 0.760 1.000 0.853
age48 0.599 0.466 0.853 1.000
```

extensions

- higher-order models:
  - AR(2): T_3 is regressed on T_2 and T_1
  - AR(p): $T_{(p+1)}$ is regressed on the p previous time points
- if we interpret a residual as part of the variable at time t that is not explained by the previous variable(s), we can use this residual as an additional predictor of the next variable:
  - without AR(1), this is a first-order moving average model MA(1)
  - with AR(1), this is an autoregressive first-order moving average model, ARMA(1)
  - typically effects of the moving averages are invariant over time (Box-Jenkins time series)

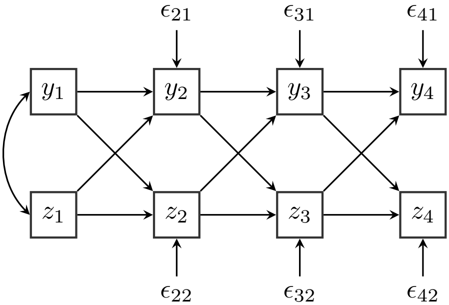
multivariate panel models

- in a multivariate panel model, we have more than one outcome, measured at (the same) t time points
- example: a bivariate panel/simplex model where Y is a measure of mathematical achievement, and Z is a measure of reading ability (4 time points: grade 3, grade 4, grade 5 and grade 6)



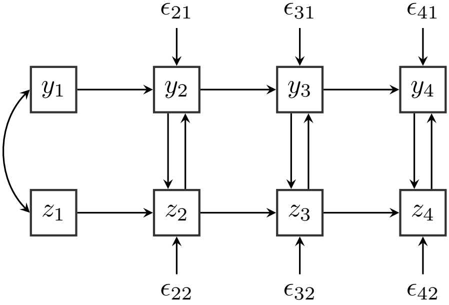
crosslagged effects

- 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?



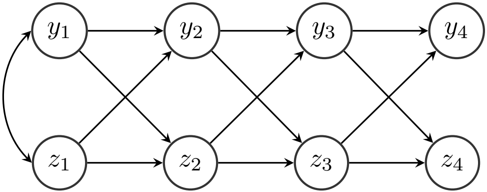
contemporaneous effects

- sometimes, the crossed effects between two variables are not lagged, but contemporaneous (exerting an effect at the same time point)
- this can be unidirectional, or reciprocal
- not everyone believes this approach is useful (in addition: often convergence issues)



panel model with latent variables

- if the 'repeated' outcomes are not directly observable, we may replace them with a latent variable with a proper measurement model
- but first, we need to establish 'measurement invariance' for the latent variables across time



- in this diagram, the observed indicators have been omitted

strengths and limitations of panel models

- panel models can be very useful for examining the relations of two (or more) variables (observed or latent) over time
- often, we are equally interested in the lack of relations over time
- panel models do not tell us anything about group level tendencies (overall increase or decrease of the scores)
- panel models do not tell us anything about individual tendencies

1.5 Growth curve models

- ‘time’ is typically considered as a continuous variable
- two components:
  - fixed effects: what is the nature of the average trend (linear, quadratic)
  - random effects: individual differences
- in addition, we may try to explain these individual differences by taking into account:
  - time-invariant covariates (age, gender, . . .)
  - time-varying covariates (measured at each time point)
- closely related to ‘mixed models’ (linear mixed models, generalized mixed models)
  - limited to balanced data
  - but we can add indirect paths and latent variables
- focus on the mean structure (not the covariance structure)

some references

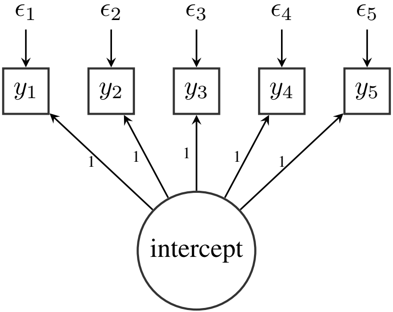
- Bollen, K.A., & Curran, P.J. (2006). *Latent curve models: A structural equation perspective.* John Wiley & Sons.
- Duncan, T.E., Duncan, S.C., & Strycker, L.A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Routledge Academic.
- Preacher, K.J., Wichman, A.L., MacCallum, R.C., & Briggs, N.E. (2008). *Latent Growth Curve Modeling.* Quantitative Applications in the Social Sciences, No. 157, Sage.

from latent variable to random effect

- a random effect is simply a latent variable with the following properties:
  - the repeated measures are the indicators of the latent variable
  - the factor loadings are fixed to a specific pattern
  - the intercepts of the observed repeated measures are fixed to zero
  - the mean/intercept of the latent variable is freely estimated
  - the (residual) variance of the latent variable is freely estimated
- typical patterns for the factor loadings:
  - by fixing all factor loadings to unity, we obtain a random intercept
  - by fixing all factor loadings to a linear scale (eg. 0, 1, 2, 3, ...) we obtain a random slope
  - by fixing all factor loadings to a quadratic scale (eg. 0, 1, 4, 9, ...), we obtain a random quadratic effect
  - ...

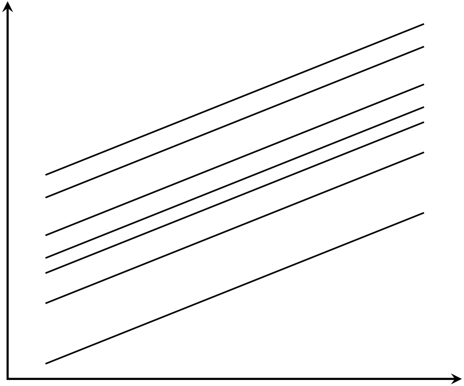
random intercept

- creating a random intercept:



random intercept only, positive linear trend

- a random-intercept-only model assumes that all individuals follow the same trend, but with a different initial point (intercept)



R code

- when using the sem() or cfa() fitting functions, you need to manually set the intercepts of the observed repeated variables to zero, and free the latent intercept:

```
> model <- '
+     # random intercept
+     int =~ 1*y1 + 1*y2 + 1*y3 + 1*y4 + 1*y5
+
+     # zero intercepts
+     y1 + y2 + y3 + y4 + y5 ~ 0*1
+
+     # free latent intercept
+     int ~ 1
+ '
```

- the growth() fitting function does this automatically (for all latent variables):

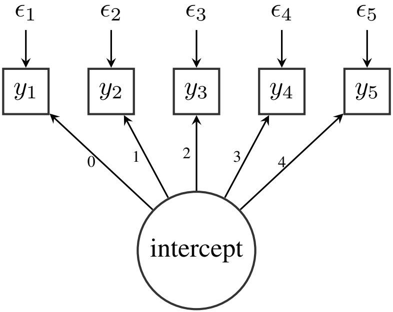
```
> model <- '
+     # random intercept
+     int =~ 1*y1 + 1*y2 + 1*y3 + 1*y4 + 1*y5
+ '
```

- when both ‘regular’ latent variables, and ‘random effects’ are used in the same model, it is perhaps better to use the lavaan() function:

```
> model <- '
+     # random intercept
+     int =~ 1*y1 + 1*y2 + 1*y3 + 1*y4 + 1*y5
+
+     # free latent intercept and variance
+     int ~ 1
+     int ~~ int
+
+     # add residual variances
+     y1 ~~ y1; y2 ~~ y2; y3 ~~ y3; y4 ~~ y4; y5 ~~ y5
+ '
```

random slope

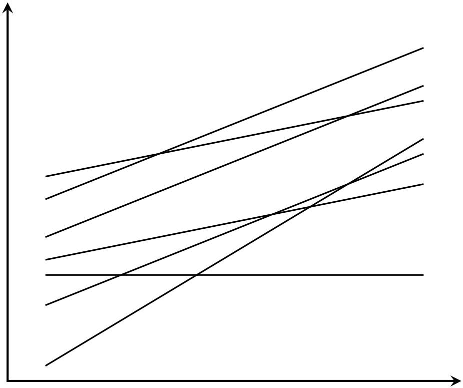
- creating a random slope:



- here, the 'reference' point is the first time point; another coding scheme (-4, -3, -2, -1, 0) treats the last time point as the reference point
- this will not affect model fit, but it will change the interpretation of the parameters

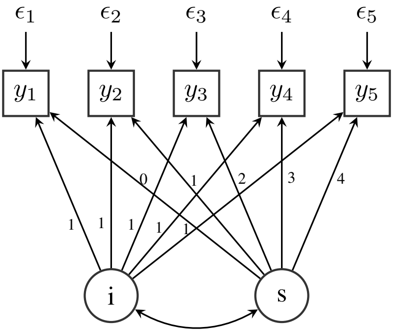
random intercept and random slope

- different intercepts, different slopes



a typical growth curve model

- random intercept and random slope



- y_t = (initial time at time 1) + (growth per unit time)*time + error
- y_t = intercept + slope*time + error

example: McCarthy Scales of Children's Abilities

- random intercept, but fixed slope

```
> model.int <- '
+     int   =~ 1*age30 + 1*age36 + 1*age42 + 1*age48
+     slope =~ 0*age30 + 1*age36 + 2*age42 + 3*age48
+
+     # intercepts (fixed effects)
+     int ~ 1
+     slope ~ 1
+
+     # random intercept
+     int ~~ int
+
+     # fixed slope
+     slope ~~ 0*slope # no variance
+     int   ~~ 0*slope # no covariance
+
+     # force same variance for all (compound symmetry)
+     age30 ~~ v1*age30
+     age36 ~~ v1*age36
+     age42 ~~ v1*age42
+     age48 ~~ v1*age48
+ '
> fit.int <- lavaan(model.int, data=MD11.5)
> summary(fit.int)
```

```
lavaan (0.5-17.700) converged normally after  56 iterations

  Number of observations                            12

  Estimator                                         ML
  Minimum Function Test Statistic               17.280
  Degrees of freedom                                10
  P-value (Chi-square)                           0.068

Parameter estimates:

  Information                                 Expected
  Standard Errors                             Standard

                   Estimate  Std.err  Z-value  P(>|z|)
Latent variables:
  int =~
    age30             1.000
    age36             1.000
    age42             1.000
    age48             1.000
  slope =~
    age30             0.000
    age36             1.000
    age42             2.000
    age48             3.000

Covariances:
```

```
  int ~~
    slope            0.000

Intercepts:
    int            103.500    3.688   28.063    0.000
    slope            3.000    0.967    3.104    0.002
    age30            0.000
    age36            0.000
    age42            0.000
    age48            0.000

Variances:
    int            123.986   56.435
    slope            0.000
    age30    (v1)   56.056   13.212
    age36    (v1)   56.056   13.212
    age42    (v1)   56.056   13.212
    age48    (v1)   56.056   13.212
```

- compare this to the output of lmer():

```
> fit <- lmer(y ~ 1  + age + (1 | subject), data=MD11.5.long, REML=FALSE)
> summary(fit)

Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: y ~ 1 + age + (1 | subject)
   Data: MD11.5.long

      AIC       BIC    logLik  deviance
 364.9290  372.4138 -178.4645  356.9290

Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 123.99   11.135
 Residual              56.06    7.487
Number of obs: 48, groups: subject, 12

Fixed effects:
            Estimate Std. Error t value
(Intercept) 103.5000     3.6881  28.063
age           3.0000     0.9666   3.104

Correlation of Fixed Effects:
    (Intr)
age -0.393
```

plausible values for the random intercepts

- the predict() function computes (estimated) individual values for the latent variables (here random effects):

```
> predict(fit.int)

          int slope
 [1,] 104.398     3
 [2,] 114.281     3
 [3,]  99.906     3
 [4,]  87.328     3
 [5,] 115.180     3
 [6,]  97.211     3
 [7,] 120.571     3
 [8,]  93.617     3
 [9,]  90.922     3
[10,]  97.211     3
[11,] 104.398     3
[12,] 116.977     3
```

- random intercept, random slope

```
> model.slope <- '
+     int   =~ 1*age30 + 1*age36 + 1*age42 + 1*age48
+     slope =~ 0*age30 + 1*age36 + 2*age42 + 3*age48
+
+     # intercepts (fixed effects)
+     int ~ 1
+     slope ~ 1
+
+     # random intercept, random slope
+     int ~~ int
+     slope ~~ slope
+     int ~~ slope
+
+     # force same variance for all (compound symmetry)
+     age30 ~~ v1*age30
+     age36 ~~ v1*age36
+     age42 ~~ v1*age42
+     age48 ~~ v1*age48
+ '
> fit.slope <- lavaan(model.slope, data=MD11.5)
> summary(fit.slope)

lavaan (0.5-17.700) converged normally after  85 iterations

  Number of observations                            12
```

```
  Estimator                                         ML
  Minimum Function Test Statistic               12.610
  Degrees of freedom                                 8
  P-value (Chi-square)                           0.126

Parameter estimates:

  Information                                 Expected
  Standard Errors                             Standard

                   Estimate  Std.err  Z-value  P(>|z|)
Latent variables:
  int =~
    age30             1.000
    age36             1.000
    age42             1.000
    age48             1.000
  slope =~
    age30             0.000
    age36             1.000
    age42             2.000
    age48             3.000

Covariances:
  int ~~
    slope           -17.682   19.168   -0.922    0.356

Intercepts:
```

```
    int               103.500    3.852   26.870    0.000
    slope               3.000    1.281    2.341    0.019
    age30               0.000
    age36               0.000
    age42               0.000
    age48               0.000

Variances:
    int               153.668   73.024
    slope              12.743    8.293
    age30     (v1)     34.817   10.051
    age36     (v1)     34.817   10.051
    age42     (v1)     34.817   10.051
    age48     (v1)     34.817   10.051


> anova(fit.int, fit.slope)

Chi Square Difference Test

          Df    AIC    BIC Chisq Chisq diff Df diff Pr(>Chisq)
fit.slope  8 364.26 367.17 12.61
fit.int   10 364.93 366.87 17.28     4.6704       2    0.09679 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- compare this to the output of lmer():

```
> fit <- lmer(y ~ 1  + age + (1 + age | subject), data=MD11.5.long, REML=FALS
> summary(fit)

Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: y ~ 1 + age + (1 + age | subject)
   Data: MD11.5.long

      AIC       BIC    logLik  deviance
 364.2587  375.4859 -176.1293  352.2587

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 153.67   12.396
          age          12.74    3.570   -0.40
 Residual              34.82    5.901
Number of obs: 48, groups: subject, 12

Fixed effects:
            Estimate Std. Error t value
(Intercept)  103.500      3.852  26.870
age            3.000      1.281   2.341

Correlation of Fixed Effects:
    (Intr)
age -0.475
```

plausible values for the random intercepts and slopes

- (estimated) individual values for both random effects:

```
> predict(fit.slope)

           int  slope
 [1,] 101.922  4.684
 [2,] 107.922  7.569
 [3,]  99.582  3.115
 [4,]  84.612  4.356
 [5,] 119.082  0.721
 [6,] 106.437 -3.360
 [7,] 124.863  0.614
 [8,]  87.851  6.579
 [9,]  92.584  1.527
[10,]  96.980  2.974
[11,] 106.168  1.840
[12,] 113.997  5.382
```

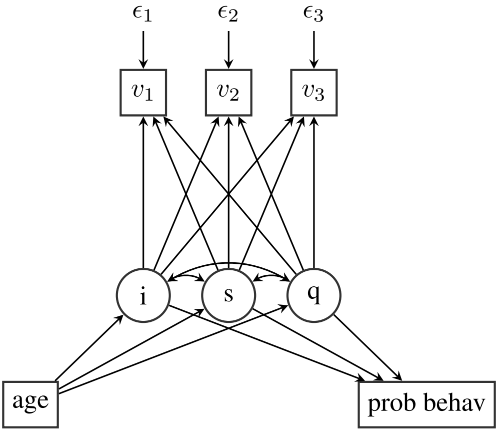
comments

- in the previous examples, we forced the residual variances to be equal, but this was just to mimic the default behavior of the linear mixed model
- in general, it is often better to leave the residual variances free
- a quadratic random effect can be added by simply adding another latent variable with quadratic factor loadings (0, 1, 4, 9); similar for cubic and higher effects
- the random effects are exogenous (no incoming arrows); we call this the *unconditional* growth curve model
- instead of fixing the linear scale (as in 0, 1, 2, 3), we may estimate some of the factor loading (as in 0, 1, ?, ?): this allows for more flexible shapes (not necessarily linear)
- custom shapes (i.e. piecewise curves) are also possible

second example (optional)

- example from Chapter 3 from the book ‘An introduction to latent variable growth curve modeling’ (Duncan et. al. 2006); sample statistics from table 3.4; model from example 3.4
- repeated measures: alcohol use at three time points (V1, V2, V3)
- possible consequence: problem behaviour
- possible predictor of growth factors: age
- we will model the time trend by a quadratic curve: this is perhaps a bit too ambitious, since we only have three time points; the price we have to pay is to fix the residual variances of the repeated measure to zero
- orthogonal coding of the intercept/slope/quadratic components

Path diagram



R code

```
> lower <- ' 0.481
+            0.401 0.539
+            0.168 0.044 0.120
+            0.311 0.343 0.516 -0.002 '
> COV3.4 <- getCov(x=lower, diag=FALSE,
+                 sds=c(7.388,  8.000,  8.043, 0.379, 0.790),
+                 names=c("v1","v2","v3","age", "probelmbehav"))
> MEAN3.4 <- c(8.265, 10.084, 10.888, 15.363, 0.026)
> model3.4 <- '
+     int  =~ 0.577*v1 + 0.577*v2 + 0.577*v3
+     linear    =~ (-0.707)*v1 + 0*v2 + 0.707*v3
+     quadratic =~ 0.408*v1 + (-0.816)*v2 + 0.408*v3
+
+     # random effects: means and variances
+     int ~ 1; int ~~ int
+     linear ~ 1; linear ~~ linear
+     quadratic ~ 1; quadratic ~~ quadratic
+
+     # fix error variances to zero
+     v1 ~~ 0*v1
+     v2 ~~ 0*v2
+     v3 ~~ 0*v3
+
+     # effect of age
+     int + linear + quadratic ~ age
+     age ~ 1
```

```
+       age ~~ age
+
+       # sequelae of change
+       probelmbehav ~ 1 + int + linear + quadratic
+       probelmbehav ~~ probelmbehav
+
+       # correlated residuals
+       int ~~ linear
+       int ~~ quadratic
+       linear ~~ quadratic
+ '
> fit3.4 <- lavaan(model3.4, sample.cov=COV3.4, sample.mean=MEAN3.4,
+                  sample.nobs=358, mimic="EQS")
> summary(fit3.4, standardized=TRUE)

lavaan (0.5-17.700) converged normally after 246 iterations

  Number of observations                           358

  Estimator                                         ML
  Minimum Function Test Statistic                2.974
  Degrees of freedom                                 1
  P-value (Chi-square)                           0.085

Parameter estimates:

  Information                                 Expected
  Standard Errors                             Standard
```

```
                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Latent variables:
  int =~
    v1                0.577                               6.299    0.853
    v2                0.577                               6.299    0.787
    v3                0.577                               6.299    0.783
  linear =~
    v1               -0.707                              -4.231   -0.573
    v2                0.000                               0.000    0.000
    v3                0.707                               4.231    0.526
  quadratic =~
    v1                0.408                               2.176    0.295
    v2               -0.816                              -4.352   -0.544
    v3                0.408                               2.176    0.271

Regressions:
  int ~
    age               3.900    1.511    2.582    0.010    0.357    0.135
  linear ~
    age              -0.515    0.835   -0.617    0.537   -0.086   -0.033
  quadratic ~
    age               1.619    0.740    2.189    0.029    0.304    0.115
  probelmbehav ~
    int               0.035    0.003   10.609    0.000    0.380    0.481
    linear            0.023    0.006    3.821    0.000    0.136    0.172
    quadratic         0.018    0.007    2.758    0.006    0.098    0.124
```

```
Covariances:
  int ~~
    linear            6.972    3.444    2.024    0.043    0.108    0.108
    quadratic        -6.614    3.053   -2.166    0.030   -0.115   -0.115
  linear ~~
    quadratic        -0.573    1.678   -0.341    0.733   -0.018   -0.018

Intercepts:
    int             -43.020   23.214   -1.853    0.064   -3.940   -3.940
    linear            9.768   12.837    0.761    0.447    1.632    1.632
    quadratic       -25.291   11.370   -2.224    0.026   -4.742   -4.742
    age              15.363    0.020  765.898    0.000   15.363   40.536
    probelmbehav     -0.596    0.066   -9.091    0.000   -0.596   -0.755
    v1                0.000                               0.000    0.000
    v2                0.000                               0.000    0.000
    v3                0.000                               0.000    0.000

Variances:
    int             117.010    8.758   13.360    0.000    0.982    0.982
    linear           35.781    2.678   13.360    0.000    0.999    0.999
    quadratic        28.072    2.101   13.360    0.000    0.987    0.987
    v1                0.000                               0.000    0.000
    v2                0.000                               0.000    0.000
    v3                0.000                               0.000    0.000
    age               0.144    0.011   13.360    0.000    0.144    1.000
    probelmbehav      0.449    0.034   13.360    0.000    0.449    0.719
```

comments

- in the model, we have not included a direct effect of age on subsequent problem behavior (hence we have 1 degree of freedom)
- age only affects the constant and the quadratic growth factors
- but all three latent growth factors are significant predictors of problem behavior
- the main conclusion of this model was that the effect of 'age' on problem behavior was mediated through its effect on the developmental parameters (the growth factors); adding a direct effect of age on problem behavior did not significantly improve the fit

autoregressive latent trajectory (ALT) models

- Bollen & Curran, in a series of papers (1999, 2000, 2001, 2004) proposed a hybrid model they called the 'autoregressive latent trajectory' (ALT) model; best reference:

  Bollen, K.A., & Curran, P.J. (2004). Autoregressive latent trajectory (ALT) models: a synthesis of two traditions. *Sociological Methods & Research*, 32(3), 336-383.

- it is a growth curve model, combined with an autoregressive structure for the repeated measures
- can be used when there is interest in both continuous underlying trajectories and time-specific influences across constructs
- the authors described this approach as 'The Best of Both Worlds'
- nevertheless, the approach has been criticized because the interpretation is not always clear

example ALT model

- combination of a growth curve model with a random intercept, a random slope, and an autoregressive structure for the repeated measures