STATISTISK ANALYS AV KOMPLEXA DATA

LONGITUDINELLA DATA

Linda Wänström

Linköpings universitet

Longitudinella data

- Tvärsnittsdata
  - Flera enheter undersöks under samma tidpunkt
  - Enheterna antas vara oberoende av varandra
- Tidsseriedata
  - En enhet undersöks under många tidpunkter
  - Observationer vid närliggande tidpunkter antas ej vara oberoende av varandra
- Longitudinella data / Paneldata
  - Flera enheter undersöks under flera tidpunkter
  - Enheterna antas vara oberoende av varandra
  - Observationer vid närliggande tidpunkter antas ej vara oberoende av varandra

Longitudinella data - exempel

- National Longitudinal Survey of Youth (NLSY79)
  - Ett slumpmässigt urval av män och kvinnor mellan 14 och 19 år samlades in 1979 i USA. De har sedan undersökts varje/varannat år sedan dess. Även kvinnornas barn har undersökts varannat år sedan 1986.
  - Exempel på variabler som har samlats in hos barnen: IQ-tester, kriminella beteenden, familjebakgrund, utbildningsnivå...
- Individual Development and Adaptation (IDA)
  - Alla som gick i trean i Örebro år 1969. De har senare följts upp i sexan och nian samt i vuxen ålder.
  - Exempel på variabler som har samlats in: Betyg, IQ, familjebakgrund, utbildningsnivå, inkomst...
- Longitudinal Individual Database (LINDA)
  - Longitudinellt register hos SCB - ett stort urval av individer i Sverige (från 1969)
  - Exempel på variabler som har samlats in: Inkomst, yrke, pendling, antal personer i hushållet...

Notation

Antag att vi har ett urval av individer - för varje individ har vi gjort upprepade mätningar på en respons. För den j:te individen kan vi skriva responsvektorn

$$\mathbf{Y}_j = \begin{bmatrix} y_{j1} \\ y_{j2} \\ \dots \\ y_{jk_j} \end{bmatrix}, j = 1, 2, ..., n$$

där n är antalet individer och k_j är antalet upprepade mätningar för individ j.

Har vi även observerat oberoende variabler kan vi skriva

$Ind(j)$	$Mätning(i)$	$Tid(t_{ij})$	y_{ij}	x_{ij1}	x_{ij2}	$\dots$	x_{ijp}
1	1	t_{11}	y_{11}	x_{111}	x_{112}	$\dots$	x_{11p}
1	2	t_{12}	y_{12}	x_{121}	x_{122}	$\dots$	x_{12p}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
1	k_1	t_{1k_1}	y_{1k_1}	x_{1k_11}	x_{1k_12}	$\dots$	x_{1k_1p}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
j	1	t_{j1}	y_{j1}	x_{j11}	x_{j12}	$\dots$	x_{j1p}
j	2	t_{j2}	y_{j2}	x_{j21}	x_{j22}	$\dots$	x_{j2p}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
j	k_j	t_{jk_j}	y_{jk_j}	x_{jk_j1}	x_{jk_j2}	$\dots$	x_{jk_jp}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
n	1	t_{n1}	y_{n1}	x_{n11}	x_{n12}	$\dots$	x_{n1p}
n	2	t_{n2}	y_{n2}	x_{n21}	x_{n22}	$\dots$	x_{n2p}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
n	k_n	t_{nk_n}	y_{nk_n}	x_{nk_n1}	x_{nk_n2}	$\dots$	x_{nk_n1}

Metoder för att analysera longitudinella data - ej bra enligt mig

- t-test för korrelerade data
  - Endast för två tidpunkter
- Change scores
  - Endast för två tidpunkter
- ANOVA/ANCOVA
  - Endast för kontinuerlig respons (som antas normalfördelad)
  - Fungerar bäst för balanserade data
- MANOVA/MANCOVA
  - En multivariat generalisering av MANOVA där vi har en vektor av responsvariabler
  - Endast för kontinuerliga respons (multivariat NF)
  - Balanserade data

Metoder för att analysera longitudinella data - bra enligt mig

- Multilevel-modell (General Linear Mixed Model)
  - Fungerar bra även om individerna är mätta vid olika tidpunkter / olika antal tidpunkter
  - Kan ha mer än två nivåer
- Latent Growth Curve Model
  - En faktoranalys-approach
  - Svårt om olika tidpunkter / antal tidpunkter
  - Fungerar bra om responsvariabeln är latent
  - Kan utvidgas till mer komplicerade faktorstrukturer

Multilevel-modell

Racdom coefficent-modell med tid som oberoende variabel

$$
\begin{aligned}
y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \epsilon_{ij} \\
\beta_{0j} = \beta_0 + u_{0j} \\
\beta_{1j} = \beta_1 + u_{1j}
\end{aligned}
$$

$$\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$$

$$\left[\begin{array}{c} u_{0j} \\ u_{1j} \end{array}\right] \sim N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u10} & \sigma_{u1}^2 \end{array}\right]\right)$$

Autokorrelation

- Observationer som ligger nära varandra i tid är ofta ej oberoende (per individ)
    - Jmfr tidsserieanalys
- Antagandet $\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$ är därför ofta inte uppfyllt vid upprepade mätningar
- Vi kan anta olika kovariansstrukturer för feltermerna

$$\epsilon_{ij} \sim N\left(0, R\right)$$

Exempel på vanliga strukturer för kovariansmatrisen

Compound symmetry

$$R = \begin{bmatrix} \sigma_e^2 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_e^2 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_e^2 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma_e^2 \end{bmatrix}$$

AR(1) (fungerar ej om olika tidsintervall)

$$R = \sigma_e^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Unstructured

$$R = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

Jämförelse av kovariansstrukturer

Modeller med små värden på $-2LL$ samt få skattade parametrar är ofta att föredra. Ett jämförelsemått som tar hänsyn till antal skattade parametrar i kovariansmatrisen (q) är AIC. Den modell med lägst AIC föredras.

$$AIC = -2LL + 2q$$

Om skattningsmetoden är REML så kan jämförelsemått baserade på -2LL endast användas då man jämför modeller med olika slumpmässiga parametrar.

Multilevel regressionsmodell med tidspolynom

Ju fler upprepade mätningar, desto större tidspolynom kan modelleras

$$
\begin{aligned}
y_{ij} &= \beta_{0j} + \beta_{1j} t_{ij} + \beta_{2j} t_{ij}^2 + \ldots + \epsilon_{ij} \\
&\quad \beta_{0j} = \beta_0 + u_{0j} \\
&\quad \beta_{1j} = \beta_1 + u_{1j} \\
&\quad \beta_{2j} = \beta_2 + u_{2j}
\end{aligned}
$$

....

$$\epsilon_{ij} \sim N(0, R)$$

$$
\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} & \sigma_{u02} \\ \sigma_{u10} & \sigma^2_{u1} & \sigma_{u12} \\ \sigma_{u2} & \sigma_{u21} & \sigma^2_{u2} \end{bmatrix} \right)
$$

Multilevel regressionsmodell med oberoende variabler

Precis som vid "vanliga" mulitlevelmodeller kan oberoende variabler läggas till på respektive nivå

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_{2j} t_{ij}^2 + \beta_3 x_{ij1} + \epsilon_{ij}$$
$$\beta_{0j} = \beta_0 + \beta_4 x_{j2} + u_{0j}$$
$$\beta_{1j} = \beta_1 + \beta_5 x_{j2} + u_{1j}$$
$$\beta_{2j} = \beta_2 + \beta_6 x_{j2} + u_{2j}$$

$$\epsilon_{ij} \sim N(0, R)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} & \sigma_{u02} \\ \sigma_{u10} & \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u2} & \sigma_{u21} & \sigma_{u2}^2 \end{bmatrix} \right)$$

Exempel: Barn från NLSY79 (de första observationerna)

Kolumnerna ger: observation, individ, familj, etnicitet, kön, födelseår, födelseordning, mors ålder vid barnets födsel, Piat Math, Piat Math åldersstandardiserat, ålder (månader)

Obs	id	idmom	race	sex	birthyear	birthorder	agemom	m	mz	a
1	201	2	3	2	1993	1	34	10	90	63
2	201	2	3	2	1993	1	34	24	100	87
3	201	2	3	2	1993	1	34	38	95	111
4	201	2	3	2	1993	1	34	55	111	136
5	202	2	3	2	1994	2	35	13	103	66
6	202	2	3	2	1994	2	35	27	102	91
7	202	2	3	2	1994	2	35	47	110	115
8	301	3	3	2	1981	1	19	27	107	85
9	301	3	3	2	1981	1	19	49	104	134
10	301	3	3	2	1981	1	19	57	102	159

Multilevel regressionsmodell ("random coefficent"-modell)

$$
\begin{aligned}
m_{ij} &= \beta_{0j} + \beta_{1j} a_{ij} + \epsilon_{ij} \\
\beta_{0j} &= \beta_0 + u_{0j} \\
\beta_{1j} &= \beta_1 + u_{1j}
\end{aligned}
$$

$$\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u10} & \sigma_{u1}^2 \end{bmatrix} \right)$$

```
proc mixed data=four covtest;
class id;
model m=a /solution ddfm=bw;
random intercept a/ subject=id type=un gcorr;
run;
```

Varning i log om G-matris

Estimated G Correlation Matrix				
Row	**Effect**	**ID CODE OF CHILD**	**Col1**	**Col2**
1	Intercept	201	1.0000	
2	a	201		1.0000

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	**id**	0	.	.	.
UN(2,1)	**id**	-0.1084	0.01113	-9.74	<.0001
UN(2,2)	**id**	0.006052	0.000210	28.82	<.0001
Residual		46.6271	0.4176	111.65	<.0001

Fit Statistics	
-2 Res Log Likelihood	241632.7
AIC (smaller is better)	241638.7
AICC (smaller is better)	241638.7
BIC (smaller is better)	241660.1

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
2	13705.46	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-12.7128	0.1440	9203	-88.28	<.0001
a	0.4353	0.001485	25E3	293.06	<.0001

Type 3 Tests of Fixed Effects

Standardiserad ålder (a)

```
proc mixed data=zfive covtest;
class id;
model m=a /solution ddfm=bw;
random intercept a/ subject=id type=un gcorr;
run;
```

Fortfarande varning i log om G-matris

Estimated G Correlation Matrix				
Row	**Effect**	**ID CODE OF CHILD**	**Col1**	**Col2**
1	Intercept	201	1.0000	1.0000
2	a	201	1.0000	1.0000

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	**id**	57.4540	1.0835	53.03	<.0001
UN(2,1)	**id**	20.3649	0.4535	44.90	<.0001
UN(2,2)	**id**	4.1227	0.3008	13.70	<.0001
Residual		49.5519	0.5165	95.94	<.0001

Fit Statistics	
-2 Res Log Likelihood	241486.7
AIC (smaller is better)	241494.7
AICC (smaller is better)	241494.7
BIC (smaller is better)	241523.2

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
3	13844.47	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	38.1548	0.08981	9203	424.83	<.0001
a	14.7553	0.04715	25E3	312.95	<.0001

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
a	1	25E3	97940.6	<.0001

Åldersstandardiserad Piat Math (mz) och standardiserad ålder (a)

```
proc mixed data=zfive covtest;
class id;
model mz=a /solution ddfm=bw;
random intercept a/ subject=id type=un gcorr;
run;
```

Ingen varning!

Estimated G Correlation Matrix				
Row	**Effect**	**ID CODE OF CHILD**	**Col1**	**Col2**
1	Intercept	201	1.0000	0.3351
2	a	201	0.3351	1.0000

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	**id**	127.18	2.2396	56.79	<.0001
UN(2,1)	**id**	12.1196	0.7614	15.92	<.0001
UN(2,2)	**id**	10.2877	0.5167	19.91	<.0001
Residual		69.1573	0.7332	94.33	<.0001

Fit Statistics	
-2 Res Log Likelihood	260404.0
AIC (smaller is better)	260412.0
AICC (smaller is better)	260412.0
BIC (smaller is better)	260440.5

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
3	15670.27	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	100.36	0.1287	9202	779.84	<.0001
a	0.2740	0.06259	25E3	4.38	<.0001

Multilevel regressionsmodell

Lägg till ett tidspolynom

$$
\begin{aligned}
mz_{ij} &= \beta_{0j} + \beta_{1j} a_{ij} + \beta_2 a_{ij}^2 + \epsilon_{ij} \\
\beta_{0j} &= \beta_0 + u_{0j} \\
\beta_{1j} &= \beta_1 + u_{1j}
\end{aligned}
$$

$$\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$$

$$\left[\begin{array}{c} u_{0j} \\ u_{1j} \end{array}\right] \sim N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u10} & \sigma_{u1}^2 \end{array}\right]\right)$$

```
proc mixed data=zfive covtest;
class id;
model mz=a a*a/solution ddfm=bw;
random intercept a/ subject=id type=un gcorr;
run;
```

Estimated G Correlation Matrix				
Row	**Effect**	**ID CODE OF CHILD**	**Col1**	**Col2**
1	Intercept	201	1.0000	0.3179
2	a	201	0.3179	1.0000

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	**id**	127.06	2.2337	56.88	<.0001
UN(2,1)	**id**	11.3388	0.7563	14.99	<.0001
UN(2,2)	**id**	10.0157	0.5085	19.69	<.0001
Residual		68.4808	0.7271	94.18	<.0001

Fit Statistics	
-2 Res Log Likelihood	260133.5
AIC (smaller is better)	260141.5
AICC (smaller is better)	260141.5
BIC (smaller is better)	260170.0

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
3	15725.64	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	101.19	0.1379	9202	734.00	<.0001
a	0.3043	0.06213	25E3	4.90	<.0001
a*a	-0.8835	0.05309	25E3	-16.64	<.0001

Multilevel regressionsmodell

Även som slumpmässig effekt

$$
\begin{aligned}
mz_{ij} &= \beta_{0j} + \beta_{1j} a_{ij} + \beta_{2j} a_{ij}^2 + \epsilon_{ij} \\
\beta_{0j} &= \beta_0 + u_{0j} \\
\beta_{1j} &= \beta_1 + u_{1j} \\
\beta_{2j} &= \beta_2 + u_{2j}
\end{aligned}
$$

$$\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$$

$$
\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} & \sigma_{u02} \\ \sigma_{u10} & \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u20} & \sigma_{u21} & \sigma_{u2}^2 \end{bmatrix} \right)
$$

```
proc mixed data=zfive covtest;
class id;
model mz=a a*a/solution ddfm=bw;
random intercept a a*a/ subject=id type=un gcorr;
run;
```

Estimated G Correlation Matrix					
Row	**Effect**	**ID CODE OF CHILD**	**Col1**	**Col2**	**Col3**
1	Intercept	201	1.0000	0.4775	-0.4310
2	a	201	0.4775	1.0000	-0.8463
3	a*a	201	-0.4310	-0.8463	1.0000

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	**id**	145.58	2.8025	51.95	<.0001
UN(2,1)	**id**	19.1642	0.8695	22.04	<.0001
UN(2,2)	**id**	11.0668	0.5091	21.74	<.0001
UN(3,1)	**id**	-12.4582	0.8777	-14.19	<.0001
UN(3,2)	**id**	-6.7442	0.3149	-21.42	<.0001
UN(3,3)	**id**	5.7387	0.4182	13.72	<.0001
Residual		62.1849	0.7451	83.46	<.0001

Fit Statistics	
-2 Res Log Likelihood	259407.2
AIC (smaller is better)	259421.2
AICC (smaller is better)	259421.3
BIC (smaller is better)	259471.1

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
6	16451.86	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	101.20	0.1447	9202	699.18	<.0001
a	0.2238	0.06209	25E3	3.60	0.0003
a*a	-0.9076	0.05748	25E3	-15.79	<.0001

Multilevel regressionsmodell

Lägg till en variabel som varierar mellan individer men ej över tid

$$
\begin{aligned}
mz_{ij} &= \beta_{0j} + \beta_{1j} a_{ij} + \beta_{2j} a_{ij}^2 + \epsilon_{ij} \\
\beta_{0j} &= \beta_0 + \beta_2 agemom + u_{0j} \\
\beta_{1j} &= \beta_1 + \beta_3 agemom + u_{1j} \\
\beta_{2j} &= \beta_2 + \beta_4 agemom + u_{2j}
\end{aligned}
$$

$$\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} & \sigma_{u02} \\ \sigma_{u10} & \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u20} & \sigma_{u21} & \sigma_{u2}^2 \end{bmatrix} \right)$$

```
proc mixed data=zfive covtest;
class id;
model mz=a a*a agemom a*agemom a*a*agemom/solution ddfm=bw
random intercept a a*a/ subject=id type=un gcorr;
run;
```

Estimated G Correlation Matrix					
Row	**Effect**	**ID CODE OF CHILD**	**Col1**	**Col2**	**Col3**
1	Intercept	201	1.0000	0.4460	-0.4108
2	a	201	0.4460	1.0000	-0.8345
3	a*a	201	-0.4108	-0.8345	1.0000

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	**id**	132.08	2.6060	50.68	<.0001
UN(2,1)	**id**	16.7024	0.8250	20.25	<.0001
UN(2,2)	**id**	10.6165	0.4999	21.24	<.0001
UN(3,1)	**id**	-11.2844	0.8564	-13.18	<.0001
UN(3,2)	**id**	-6.4993	0.3104	-20.94	<.0001
UN(3,3)	**id**	5.7142	0.4202	13.60	<.0001
Residual		62.1419	0.7445	83.47	<.0001

Fit Statistics	
-2 Res Log Likelihood	258676.5
AIC (smaller is better)	258690.5
AICC (smaller is better)	258690.5
BIC (smaller is better)	258740.4

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
6	15154.37	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	101.26	0.1396	9201	725.57	<.0001
a	0.2882	0.06223	25E3	4.63	<.0001
a*a	-0.8617	0.05924	25E3	-14.55	<.0001
agemom	3.5538	0.1350	9201	26.32	<.0001
a*agemom	0.6502	0.06471	25E3	10.05	<.0001
a*a*agemom	-0.3284	0.05533	25E3	-5.94	<.0001

Struktur för R: compound symmetry

$$mz_{ij} = \beta_0 + \beta_1 a_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, R)$$

där, för en individ med fyra tidpunkter, $R = \begin{bmatrix} \sigma_e^2 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_e^2 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_e^2 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma_e^2 \end{bmatrix}$

```
proc mixed data=zfive covtest;
class id wave;
model mz=a /solution ddfm=bw;
repeated wave/type=cs subject=id r;
run;
```

Estimated R Matrix for id 201				
Row	**Col1**	**Col2**	**Col3**	**Col4**
1	205.25	124.98	124.98	124.98
2	124.98	205.25	124.98	124.98
3	124.98	124.98	205.25	124.98
4	124.98	124.98	124.98	205.25

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
CS	**id**	124.98	2.2200	56.30	<.0001
Residual		80.2668	0.7218	111.21	<.0001

Fit Statistics

-2 Res Log Likelihood	261232.7
AIC (smaller is better)	261236.7
AICC (smaller is better)	261236.7
BIC (smaller is better)	261251.0

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	14841.51	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	100.26	0.1281	9202	782.86	<.0001
a	0.2888	0.05345	25E3	5.40	<.0001

Kovariansstruktur för olika tidsintervall

SP(POW) (Spatial Power Law)

$mz_{ij} = \beta_0 + \beta_1 a_{ij} + \epsilon_{ij}$

$\epsilon_{ij} \sim N(0, R)$

AR(1) går inte att använda vid olika tidsintervall. En generalisering av AR(1) för olika tidsintervall är SP(POW). Kovariansen mellan två mätningar vid tidpunkt T1 och T2 modelleras som:

$$Cov\left[Y_{t1}, Y_{t2}\right] = \sigma^2 \rho^{|t1-t2|}$$

```
proc mixed data=zfive covtest;
class id wave;
model mz=a /solution ddfm=bw;
repeated wave/type=sp(pow)(a) subject=id r;
run;
```

Estimated R Matrix for id 201				
Row	**Col1**	**Col2**	**Col3**	**Col4**
1	206.01	136.35	90.2442	58.7113
2	136.35	206.01	136.35	88.7056
3	90.2442	136.35	206.01	134.02
4	58.7113	88.7056	134.02	206.01

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
SP(POW)	**id**	0.5580	0.004752	117.44	<.0001
Residual		206.01	2.0968	98.25	<.0001

Fit Statistics	
-2 Res Log Likelihood	262431.8
AIC (smaller is better)	262435.8
AICC (smaller is better)	262435.8
BIC (smaller is better)	262450.1

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
1	13642.45	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	100.03	0.1192	9202	839.42	<.0001
a	0.06059	0.07801	25E3	0.78	0.4374

$$
\begin{aligned}
mz_{ij} &= \beta_{0j} + \beta_{1j} a_{ij} + \beta_{2j} a_{ij}^2 + \epsilon_{ij} \\
\beta_{0j} &= \beta_0 + \beta_2 agemom + u_{0j} \\
\beta_{1j} &= \beta_1 + \beta_3 agemom + u_{1j} \\
\beta_{2j} &= \beta_2 + \beta_4 agemom + u_{2j}
\end{aligned}
$$

$$\epsilon_{ij} \sim N(0, R)$$

$$
\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} & \sigma_{u02} \\ \sigma_{u10} & \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u20} & \sigma_{u21} & \sigma_{u2}^2 \end{bmatrix} \right)
$$

```
proc mixed data=zfive covtest;
class id wave;
model mz=a a*a agemom a*agemom a*a*agemom/solution ddfm=bw
random intercept a a*a/ subject=id type=un gcorr;
repeated wave/type=cs subject=id r;
run;
```

Estimated R Matrix for id 201

Row	Col1	Col2	Col3	Col4
1	61.3632	-0.7750	-0.7750	-0.7750
2	-0.7750	61.3632	-0.7750	-0.7750
3	-0.7750	-0.7750	61.3632	-0.7750
4	-0.7750	-0.7750	-0.7750	61.3632

Estimated G Correlation Matrix

Row	Effect	ID CODE OF CHILD	Col1	Col2	Col3
1	Intercept	201	1.0000	0.4447	-0.4095
2	a	201	0.4447	1.0000	-0.8342
3	a*a	201	-0.4095	-0.8342	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	id	132.86	2.6045	51.01	<.0001
UN(2,1)	id	16.7032	0.8250	20.25	<.0001
UN(2,2)	id	10.6187	0.5000	21.24	<.0001
UN(3,1)	id	-11.2856	0.8564	-13.18	<.0001
UN(3,2)	id	-6.4995	0.3104	-20.94	<.0001
UN(3,3)	id	5.7162	0.4204	13.60	<.0001
CS	id	-0.7750	0.009286	-83.46	<.0001
Residual		62.1383	0.7445	83.46	<.0001

Fit Statistics

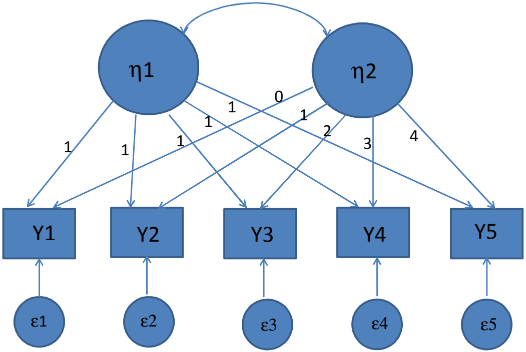
-2 Res Log Likelihood	258676.5
AIC (Smaller is Better)	258692.5
AICC (Smaller is Better)	258692.5
BIC (Smaller is Better)	258749.5

- Antagande: Korrelationer mellan ett antal observerade variabler uppstår pga bakomliggande, latenta faktorer
- Antal faktorer kan bestämmas från PC-analys
- Observerade variablers laddningar kan skattas med iterativ PC-analys

Konfirmativ Faktoranalys

- "Latent growth curve-modeller" (LGC) härstammar från konfirmativ faktoranalys
- En faktorstruktur antas och testas
- Antagande vid LGC-analys: Latenta faktorer (intercept, slope) påverkar värdena på responsvariabeln under de olika tidpunkterna.

Exempel: LGC-modell som antar linjär utveckling för fem tidpunkter



LGC-modell

Om vi har p tidpunkter och m faktorer (intercept, lutning för tid, lutning för tid i kvadrat etc.) kan vi skriva modellen:

$$y = \tau + \Lambda\eta + \varepsilon$$

där y är en $p \times 1$ vektor av observationer, τ är en $p \times 1$ vektor av intercept (ofta fixerad till 0), η är en $m \times 1$ vektor av faktorer, Λ är en $p \times m$ matris med laddningar (ofta fixerade för att definiera intercept och olika typer av lutningar), och ε är en $p \times 1$ vektor av unika faktorer som antas normalfördelade. Faktorerna kan uttryckas som avvikelser från faktormedelvärden:

$$\eta = \mu_\eta + \zeta$$

där μ_η är en $m \times 1$ vektor av medelvärden och ζ är en $m \times 1$ vektor av residualer.

LGC-modell

Variansen för y kan uttryckas som

$$\Sigma = \Lambda \Psi \Lambda' + \Theta_\epsilon$$

där Σ är en $p \times p$ kovariansmatris, Ψ är en $m \times m$ kovariansmatris för faktorer, och Θ_ϵ är en $p \times p$ kovarianmatris för unika faktorer.
Förväntade värdet för y kan uttryckas som

$$\mu_y = \tau + \Lambda \mu_\eta$$

Exempel forts

För exemplet ovan ser vektorerna och matriserna ut som följer:

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \qquad \Psi = \begin{bmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{bmatrix}$$

$$\mu_\eta = \begin{bmatrix} \mu_{\eta_1} \\ \mu_{\eta_2} \end{bmatrix}, \qquad \Theta_\epsilon = \begin{bmatrix} \theta_\epsilon & & & & \\ 0 & \theta_\epsilon & & & \\ 0 & 0 & \theta_\epsilon & & \\ 0 & 0 & 0 & \theta_\epsilon & \\ 0 & 0 & 0 & 0 & \theta_\epsilon \end{bmatrix}$$

Skattning av de okända parametrarna

Okända parametrar i modellen är medelvärden, varianser och kovarianser för faktorer, och varianser för unika faktorer

- Samma skattningsmetoder som vid konfirmativ faktoranalys
- Elementen i observerad kovariansmatris (S) och medelvärden för alla p variabler uttrycks som funktioner av de okända parametrarna
- Skattningarna är de värden på parametrarna som reproducerar de observerade kovarianserna och varianserna (implicerad kovariansmatris Σ) och medelvärdena så bra som möjligt
- ML-skattningar: en funktion av avvikelser mellan observerade kovariansmatrisen (S) och medelvärden ($\bar{y}$) samt implicerade kovariansmatrisen (Σ) och medelvärden (μ) minimeras

$$F_{ML} = Tr(S\Sigma^{-1}) - p + \ln(|\Sigma|) - \ln(|S|) + (\bar{y} - \mu)'\Sigma^{-1}(\bar{y} - \mu)$$

- Modellen måste vara "identifierad": antal kända bitar information (observerade varianser, kovarianser, medelvärden) måste vara minst lika många som antal okända parametrar

Test för modellen samt för dess parametrar

- Under nollhypotesen att modellen passar data, är $\chi^2 = (N-1)F$ chitvåfördelad med *df* frihetsgrader
  - *df* är antal delar information (antal observerade kovarianser, varianser och medelvärden) minus antal skattade parametrar
  - N är stickprovsstorlek
- För stora stickprov ($N > 200$) förkastas ofta nollhypotesen även om modellen har en bra anpassning till data. Anpassningsmått används därför oftast i stället.
- Test för parametrar: t-test / z-test
- Likelihood ratiotest (Chisquare difference test) /jämförelse av anpassningingsmått
  - Överkurs för denna kurs

Anpassningsmått

- CFI (Comparative Fit Index): $$CFI = \frac{(\chi^2 - df)_{Noll} - (\chi^2 - df)_{Mod}}{(\chi^2 - df)_{Noll}}$$
  - df =frihetsgrader, $Noll$ =nollmodell (alla korrelationer mellan variabler antas vara noll), Mod =skattade modellen
  - Okej anpassning: $CFI > 0.9$ (bra anpassning: $CFI > 0.95$)
  - Ej rimlig att beräkna om $RMSEA_{Noll} < 0.158$ (dvs RMSEA för nollmodellen)
- RMSEA (Root Mean Squared Error of Approximation): $$RMSEA = \frac{\sqrt{\chi^2 - df}}{\sqrt{df(N-1)}}$$
  - Mått på skillnad mellan observerad och implicerad matris
  - Okej anpassning: $RMSEA < 0.10$ (bra anpassning: $RMSEA < 0.05$)

Exempel: Ökar självsäkerhet över tid för individer som deltar i ett träningsprogram?

Självsäkerhet är mätt under 5 tidpunkter (lika intervall) för 16 individer som deltar i programmet

```
data growth;
   input y1 y2 y3 y4 y5;
   datalines;
17.6  21.4  25.6  32.1  37.7
13.2  14.3  18.9  20.3  25.4
11.6  13.5  17.4  22.1  39.6
10.7  11.1  13.2  18.2  21.4
18.7  23.7  28.6  31.5  34.0
18.3  19.2  20.5  23.2  25.9
 9.2  13.5  17.8  19.2  21.1
18.3  23.5  27.9  30.2  34.6
11.2  15.6  20.8  22.7  30.4
17.0  22.9  26.9  31.9  35.6
10.4  13.6  18.0  25.6  29.3
17.7  19.0  22.5  28.5  30.7
14.5  19.4  21.1  28.8  31.5
20.0  21.4  28.9  30.2  35.6
14.6  19.3  21.7  28.5  32.0
11.7  15.2  19.1  23.7  28.7
;
```

SAS

```
proc calis method=ml data=growth;
   lineqs
      y1 = 0. * Intercept + f_int                + e1,
      y2 = 0. * Intercept + f_int  +  1 * f_slp + e2,
      y3 = 0. * Intercept + f_int  +  2 * f_slp + e3,
      y4 = 0. * Intercept + f_int  +  3 * f_slp + e4,
      y5 = 0. * Intercept + f_int  +  4 * f_slp + e5;
   std
      f_int f_slp,
      e1-e5 = 5 * evar;
   mean
      f_int f_slp;
   cov
      f_int f_slp;
   run;
```

Variables in the Model

Manifest	y1 y2 y3 y4 y5
Latent	
Manifest	
Latent	f_int f_slp
Error	e1 e2 e3 e4 e5

Number of Endogenous Variables = 5
Number of Exogenous Variables = 7

Variable	Mean	Std Dev
y1	14.66875	3.56459
y2	17.91250	4.08736
y3	21.80625	4.60036
y4	26.04375	4.73863
y5	30.84375	5.41701

Fit Summary

Number of Observations	16
Number of Variables	5
Number of Moments	20
Number of Parameters	6
Number of Active Constraints	0
Baseline Model Function Value	6.6454
Baseline Model Chi-Square	99.6809
Baseline Model Chi-Square DF	10
Pr > Baseline Model Chi-Square	<.0001
Fit Function	2.0954
Chi-Square	31.4310
Chi-Square DF	14
Pr > Chi-Square	0.0048
Z-Test of Wilson & Hilferty	2.5819
Hoelter Critical N	12
Root Mean Square Residual (RMR)	1.9062
Standardized RMR (SRMR)	0.1205
Goodness of Fit Index (GFI)	0.9204

Parsimony Index	**Adjusted GFI (AGFI)**	0.8863
	Parsimonious GFI	1.2885
	RMSEA Estimate	0.2881
	RMSEA Lower 90% Confidence Limit	0.1525
	RMSEA Upper 90% Confidence Limit	0.4236
	Probability of Close Fit	0.0069
	Akaike Information Criterion	43.4310
	Bozdogan CAIC	54.0665
	Schwarz Bayesian Criterion	48.0665
	McDonald Centrality	0.5800
Incremental Index	**Bentler Comparative Fit Index**	0.8056
	Bentler-Bonett NFI	0.6847
	Bentler-Bonett Non-normed Index	0.8612
	Bollen Normed Index Rho1	0.7748
	Bollen Non-normed Index Delta2	0.7966
	James et al. Parsimonious NFI	0.9586

Estimates for Variances of Exogenous Variables

Variable Type	Variable	Parameter	Estimate	Standard Error	t Value
Latent	**f_int**	_Parm1	13.89140	5.81540	2.38873
	f_slp	_Parm2	0.80742	0.42198	1.91342
Error	**e1**	evar	3.32185	0.70031	4.74342
	e2	evar	3.32185	0.70031	4.74342
	e3	evar	3.32185	0.70031	4.74342
	e4	evar	3.32185	0.70031	4.74342
	e5	evar	3.32185	0.70031	4.74342

Covariances Among Exogenous Variables

Var1	Var2	Parameter	Estimate	Standard Error	t Value
f_int	**f_slp**	_Parm3	-0.35281	1.13815	-0.30998

Mean Parameters

Variable Type	Variable	Parameter	Estimate	Standard Error	t Value
Latent	**f_int**	_Parm4	14.15875	1.02906	13.75890
	f_slp	_Parm5	4.04813	0.27563	14.68665

Squared Multiple Correlations

Variable	Error Variance	Total Variance	R-Square
y1	3.32185	17.21324	0.8070
y2	3.32185	17.31505	0.8082
y3	3.32185	19.03169	0.8255
y4	3.32185	22.36317	0.8515
y5	3.32185	27.30948	0.8784

Standardized Results for Covariances Among Exogenous Variables

Var1	Var2	Parameter	Estimate	Standard Error	t Value
f_int	**f_slp**	_Parm3	-0.10535	0.32466	-0.32448

Modellen har för dålig anpassning, och parametrarna ska därför inte tolkas

$\chi^2(14) = 31.43, p = .005$
$CFI = 0.806,\ RMSEA = 0.288(90\%ci : 0.152; 0.424)$
Skattad kovariansmatris och medelvärdesvektor för de latenta variablerna:

$$\widehat{\Psi} = \begin{bmatrix} 13.891 & \\ -0.353 & 0.807 \end{bmatrix}, \widehat{\mu_\eta} = \begin{bmatrix} 14.159 \\ 4.048 \end{bmatrix}$$

Skattad kovariansmatris för fletermerna

$$\widehat{\Theta_\epsilon} = \begin{bmatrix} 3.322 & & & & \\ 0 & 3.322 & & & \\ 0 & 0 & 3.322 & & \\ 0 & 0 & 0 & 3.322 & \\ 0 & 0 & 0 & 0 & 3.322 \end{bmatrix}$$

Kan vi få en bättre anpassning med en kvadratisk term?

```
proc calis method=ml data=growth;
   lineqs
      y1 = 0. * Intercept + f_int                + e1,
      y2 = 0. * Intercept + f_int  +  1 * f_slp +  1 * f_qdr + e2,
      y3 = 0. * Intercept + f_int  +  2 * f_slp +  4 * f_qdr + e3,
      y4 = 0. * Intercept + f_int  +  3 * f_slp +  9 * f_qdr + e4,
      y5 = 0. * Intercept + f_int  +  4 * f_slp +  16 * f_qdr + e5;
   std
      f_int f_slp f_qdr,
      e1-e5 = 5 * evar;
   mean
      f_int f_slp f_qdr;
   cov
      f_int f_slp f_qdr;
   run;
```

Fit Summary

Number of Observations	16
Number of Variables	5
Number of Moments	20
Number of Parameters	10
Number of Active Constraints	0
Baseline Model Function Value	6.6454
Baseline Model Chi-Square	99.6809
Baseline Model Chi-Square DF	10
Pr > Baseline Model Chi-Square	<.0001
Fit Function	0.7681
Chi-Square	11.5209
Chi-Square DF	10
Pr > Chi-Square	0.3184
Z-Test of Wilson & Hilferty	0.4732
Hoelter Critical N	24
Root Mean Square Residual (RMR)	0.6902
Standardized RMR (SRMR)	0.0337
Goodness of Fit Index (GFI)	0.9721

Parsimony Index	**Adjusted GFI (AGFI)**	0.9442
	Parsimonious GFI	0.9721
	RMSEA Estimate	0.1007
	RMSEA Lower 90% Confidence Limit	0.0000
	RMSEA Upper 90% Confidence Limit	0.3078
	Probability of Close Fit	0.3495
	Akaike Information Criterion	31.5209
	Bozdogan CAIC	49.2468
	Schwarz Bayesian Criterion	39.2468
	McDonald Centrality	0.9536
Incremental Index	**Bentler Comparative Fit Index**	0.9830
	Bentler-Bonett NFI	0.8844
	Bentler-Bonett Non-normed Index	0.9830
	Bollen Normed Index Rho1	0.8844
	Bollen Non-normed Index Delta2	0.9830
	James et al. Parsimonious NFI	0.8844

Estimates for Variances of Exogenous Variables

Variable Type	Variable	Estimate	Standard Error	t Value
Latent	**f_int**	10.70646	4.48233	2.38859
	f_slp	2.77431	1.87557	1.47919
	f_qdr	0.28273	0.15140	1.86745
Error	**e1**	1.71800	0.44358	3.87298
	e2	1.71800	0.44358	3.87298
	e3	1.71800	0.44358	3.87298
	e4	1.71800	0.44358	3.87298
	e5	1.71800	0.44358	3.87298

Covariances Among Exogenous Variables

Var1	Var2	Estimate	Standard Error	t Value
f_int	f_slp	3.02204	2.07637	1.45545
f_int	f_qdr	-0.75408	0.59312	-1.27138
f_slp	f_qdr	-0.79128	0.50829	-1.55675

Mean Parameters

Variable Type	Variable	Estimate	Standard Error	t Value
Latent	f_int	14.65250	0.90289	16.22847
	f_slp	3.06062	0.57210	5.34977
	f_qdr	0.24688	0.16441	1.50160

Squared Multiple Correlations

Variable	Error Variance	Total Variance	R-Square
y1	1.71800	12.42445	0.8617
y2	1.71800	18.43487	0.9068
y3	1.71800	21.44049	0.9199
y4	1.71800	22.12437	0.9223
y5	1.71800	27.95518	0.9385

Modellen passar bättre och vi kan tolka parameterskattningarna

$\chi^2(10) = 11.521, p = 0.318$
$CFI = 0.983,\ RMSEA = 0.101(90\%ci : 0.000; 0.308)$
Skattad kovariansmatris och medelvärdesvektor för de latenta variablerna (skattningen för kvatrattermen är ej signifikant):

$$\widehat{\Psi} = \begin{bmatrix} 10.706 & & \\ 3.022 & 2.774 & \\ -0.754 & -0.791 & 0.283 \end{bmatrix}, \widehat{\mu_\eta} = \begin{bmatrix} 14.652 \\ 3.061 \\ 0.247 \end{bmatrix}$$

Skattad kovariansmatris för feltermerna

$$\widehat{\Theta_\epsilon} = \begin{bmatrix} 1.718 & & & & \\ 0 & 1.718 & & & \\ 0 & 0 & 1.718 & & \\ 0 & 0 & 0 & 1.718 & \\ 0 & 0 & 0 & 0 & 1.718 \end{bmatrix}$$

Ökar självsäkerheten över tid för individer som deltar i träningsprogrammet?

En latent tillväxtmodell med kvadratisk utveckling skattades och anpassningen var okej ($CFI = 0.983$, $RMSEA = 0.101$ $(90\%ci : 0.000; 0.308)$). Den initiala självsäkerheten skattades till 14.653, med en varians på 10.706. Den linjära ökningen per tidpunkt skattades till 3.061 med en varians på 2.774, och den kvadratiska ökningen skattades till 0.247 med en varians på 0.283. Den kvadratiska ökningen var ej signifikant (vilket kan bero på för liten stickprovsstorlek). Detsamma gällde varianserna för den linjära och kvadratiska utvecklingen. Den totala variansen i självsäkerhet mätt under de fem tidpunkterna förklaras till mellan 86% (vid första tidpunkten) och 94% (vid femte tidpunkten) av modellen. Sammanfattningsvis så ökar självsäkerheten för individer som deltar i träningsprogrammet, och det finns en tendens till att ökningen blir större över tid.

Multilevel-modeller eller LGC-modeller för longitudinella data?

- Om man har balanserade data (dvs alla individer är mätta vid alla tidpunkter) kan bägge metoderna användas
- Om man inte har balanserade data (eller bortfall) är det lättare med multilevel-modeller
- Oberoende variabler kan användas vid bägge metoder
- Kovariansstruktur för feltermer (unika faktorer) kan användas vid bägge metoder
- Om någon/några variabler är latent, och mätt med flera indikatorer (tex 3 IQ-test som antas mäta intelligens vid varje tidpunkt) kan man använda LGC

Multilevel-modeller eller LGC-modeller för longitudinella data?

- Om man har mer komplicerade modeller, så som att en variabel X_1 påverkar en variabel X_2 som i sin tur antas påverka interceptet och lutningen, kan man använda LGC
- Om man har mer komplicerade modeller, så som att en utvecklingskurva (intercept och lutning för exempelvis intelligensutveckling) antas påverka en annan utvecklingskurva (intercept och lutning för exempelvis betygsutveckling) kan man använda LGC
- Multilevel-modeller används oftare än LGC-modeller, men användningen av LGC-modeller ökar mer och mer!