STATISTISK ANALYS AV KOMPLEXA DATA
HIERARKISKA DATA

Linda Wänström

Linköpings universitet

Regressionsmodell för icke-hierarkiska data (repetition)

Antag att vi observerar n oberoende värden $y_1, ..., y_n$ och vill förklara dem med n värden på p förklarande variabler $x_{11}, ...x_{1p}, x_{21}, ..., x_{2p}, x_{n1}, ..., x_{np}$. Standardregressionsmodellen är, för individ i,

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$$

där $\epsilon_i \sim N(0, \sigma_\epsilon^2)$.
I matrisform kan modellen skrivas som

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

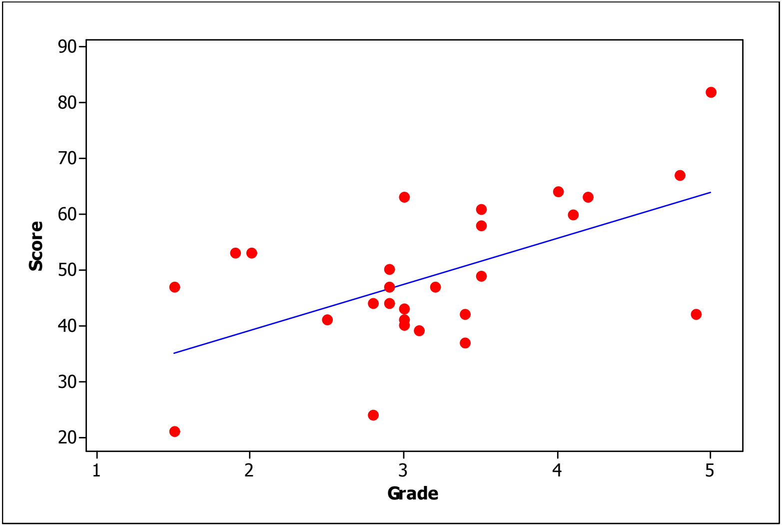
där $\mathbf{e} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$.

Hierarkiska data

- Hierarkiska data, "Clustered" data, fler-nivå data, grupperade data
- **Enheter** grupperade i olika **nivåer**
  - Vi har ett urval av skolor - inom varje skola har vi ett antal klasser - inom varje klass har vi ett urval av elever
    - Nivå 1: elever
    - Nivå 2: klasser
    - Nivå 3: skolor
  - Vi har ett urval av familjer - inom varje familj undersöker vi alla/några barn
    - Nivå 1: barn
    - Nivå 2: familjer

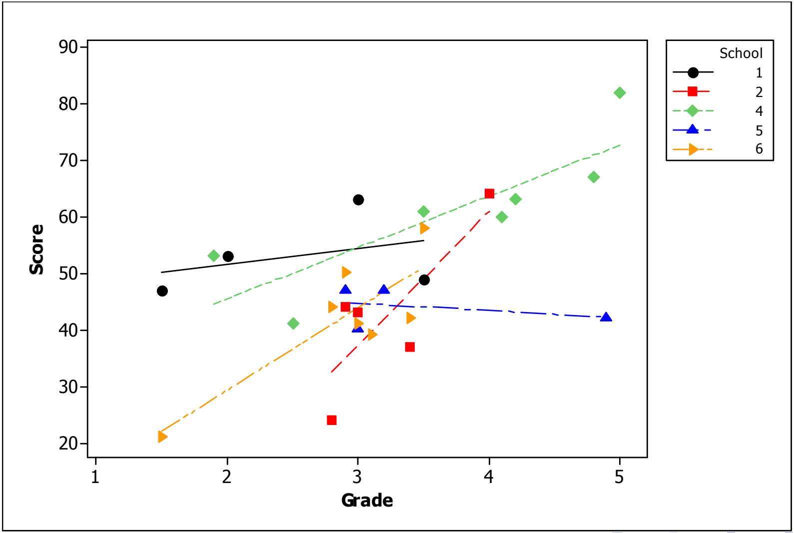
Exempel: Samband mellan medelbetyg och poäng på prov

Inritad regressionslinje



Exempel: Samband mellan medelbetyg och poäng på prov

Inritad regressionslinje för varje skola



Varför ska vi ta hänsyn till den hierarkiska strukturen på data?

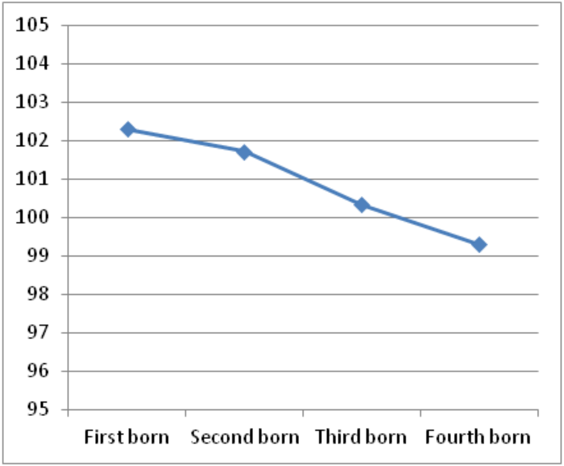
- Vi kan väl bara använda den lägsta nivån (tex elev) som enhet och använda vanliga analysmetoder?
- Eller - vi kan väl använda den högre nivån (tex klass / skola) som enhet och använda vanliga metoder?
- Individer (observationer) inom en grupp tenderar att vara lika varandra
  - Familjer: syskon (samma föräldrar, samma miljöpåverkan)
  - Skolor: klasser: elever (påverkar varandra, mer lika från början?)
- Ignorerar vi grupperingar kan vi missa viktiga gruppeffekter
- Ignorerar vi korrelationen mellan observationer inom grupper får vi missvisande skattningar på varianser

Exempel: Studie om läsinlärning i skolan (Bennet, 1976)

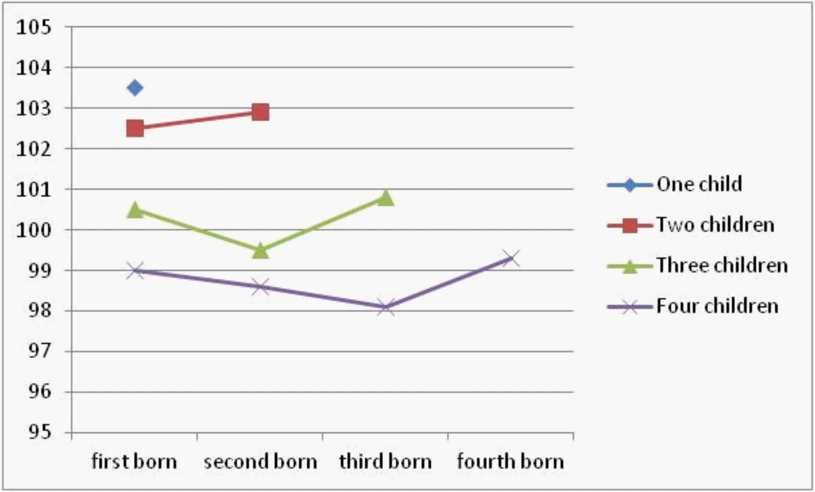
- Slutsats: Barn som lär sig läsa enligt den "traditionella tekniken" lär sig fortare än de som inte gör det
  - Multipel regressionsmodell med elever som enheter (inga grupperingar)
  - Statistisk signifikant skillnad
- Aitkin et el (1981) analyserade om datamaterialet och tog hänsyn till att eleverna var grupperade i klasser
  - Inte längre statistiskt signifikant skillnad

Exempel: Mean PIAT Scores

Data från Rodgers, J.L., Cleveland, H.H., van den Oord, E., & Rowe, D.C. (2000). Resolving the debate over birth order, family size and intelligence. American Psychologist, 55, 599-612.



Mean Piat Score uppdelat på familjestorlek



Fixed effects modeller

- Definiera dummyvariabler för att skatta skillnader mellan skolor, klasser, familjer...
- Detta fungerar om vi har ett mindre antal skolor, klasser, familjer... och endast vill generalisera till dem
- Vill vi generalisera resultaten till ett större antal skolor, klasser, familjer - använd inte fixed effects modeller
- Fungerar inte heller bra om vi har få observationer (tex elever) i några av grupperna (tex klasser)

General linear mixed model

- Multilevel model (Goldstein)
- Hierarchical linear model (Raudenbush & Bryk)
- Variance component model (Longford)
- Random coefficient model (De Leeuw & Kreft, Longford)
- Random effects model, mixed effects model etc...

Notation

Antag att vi har ett urval av klasser - inom varje klass har vi ett urval av elever där vi har observerat en respons. För den j:te klassen 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 klasser och k_j är antalet elever i klass j.

Har vi även observerat oberoende variabler kan vi skriva

$Klass(j)$	$Elev(i)$	y_{ij}	x_{ij1}	x_{ij2}	$\ldots$	x_{ijp}
1	1	y_{11}	x_{111}	x_{112}	$\ldots$	x_{11p}
1	2	y_{21}	x_{211}	x_{212}	$\ldots$	x_{21p}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
1	k_1	$y_{k_1 1}$	$x_{k_1 11}$	$x_{k_1 12}$	$\ldots$	$x_{k_1 1p}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
j	1	y_{1j}	x_{1j1}	x_{1j2}	$\ldots$	x_{1jp}
j	2	y_{2j}	x_{2j1}	x_{2j2}	$\ldots$	x_{2jp}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
j	k_j	$y_{k_j j}$	$x_{k_j j1}$	$x_{k_j j2}$	$\ldots$	$x_{k_j jp}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
n	1	y_{1n}	x_{1n1}	x_{1n2}	$\ldots$	x_{1np}
n	2	y_{2n}	x_{2n1}	x_{2n2}	$\ldots$	x_{2np}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
n	k_n	$y_{k_n n}$	$x_{k_n n1}$	$x_{k_n n2}$	$\ldots$	$x_{k_n np}$

General linear mixed model (GLMM)

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

$$\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{R}\right)$$

$$\mathbf{u} \sim N\left(\mathbf{0}, \mathbf{G}\right)$$

$$Cov\left[\mathbf{u}, \mathbf{e}\right] = \mathbf{0}$$

Modellens parametrar består av fixa effekter i vektorn β och alla okända parametrar i matriserna $\mathbf{G}$ och $\mathbf{R}$. De slumpmässiga effekterna $\mathbf{u}$ och $\mathbf{e}$ är inte parametrar utan består av slumpmässiga variabler.

General linear mixed model

Betingade fördelningen $\mathbf{Y}|\mathbf{u}$ och marginalfödelningen för $\mathbf{Y}$ är

$$\mathbf{Y}|\mathbf{u} \sim N\left(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}, \mathbf{R}\right)$$

$$\mathbf{Y} \sim N\left(\mathbf{X}\beta, \mathbf{V}\right)$$

$$\mathbf{V} = Var\left[\mathbf{Y}\right] = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

Skattning av parametrar

$-2LL$ för GLMM är

$$-2l(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{y}) = \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + c$$

ML-skattningen av fix effekt-parametrarna är

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{y}$$

Skattning av parametrar

- Vid känd varians **V** är GLS-skattning av fixa parametrar

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$
$$\mathbf{V}(\widehat{\beta}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$$

- Vanliga startvärden för fixa parametrar: OLS
- Vektor med residualer skapas, och används för att uttryckas som funktion av varianser för slumptermerna och dessa skattas (krävs att dessa är normalfördelade)
- Skattad **V** används för att ta fram nya GLS-skattningar, och nya varianser för slumptermerna skattas
- osv till konvergens
- Om slumptermerna (residualerna) är normalfördelade är skattningarna ML-skattningar
- Skattningarna för varianserna för slumptermerna är inte VVR (tar inte hänsyn till samplingvariation av fixa parametrar)
- En modifikation ger REML - VVR skattningar (används i SAS)

$$y_{ij} = \beta_{0j} + \epsilon_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$

$$y_{ij} = \beta_0 + u_{0j} + \epsilon_{ij}$$
$$y_{ij} = (\beta_0) + (u_{0j} + \epsilon_{ij})$$

$$\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$$
$$u_{0j} \sim N\left(0, \sigma_{u0}^2\right)$$

En fix (intercept) och två slumpmässiga parametrar (två varianser) att skatta

Multilevelmodell: Varianskomponentmodell

$$y_{ij} = \beta_0 + (u_{0j} + \epsilon_{ij})$$

$$\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$$
$$u_{0j} \sim N\left(0, \sigma_{u0}^2\right)$$

$$var(y_{ij}) = var(u_{0j} + \epsilon_{ij}) = \sigma_{u0}^2 + \sigma_e^2$$
$$cov(y_{1j}, y_{2j}) = cov(u_{0j} + \epsilon_{1j}, u_{0j} + \epsilon_{2j}) = \sigma_{u0}^2$$

Intraklass-korrelation

Varianskomponentmodell

$$\rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_e}$$

- Andelen av variationen som är mellan grupper (klasser)
- Om den är större än noll bör vi inte använda vanliga metoder (som tex OLS) som inte tar hänsyn till varianser på flera nivåer

Multilevelmodell: "random coefficent"-modell

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

$$
\begin{aligned}
y_{ij} &= \beta_0 + u_{0j} + \beta_1 x_{ij} + u_{1j}x_{ij} + \epsilon_{ij} \\
y_{ij} &= (\beta_0 + \beta_1 x_{ij}) + (u_{0j} + u_{1j}x_{ij} + \epsilon_{ij})
\end{aligned}
$$

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

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

Två fixa (intercept och slope) och fyra slumpmässiga parametrar (tre varianser och en kovarians) att skatta

Hypotestest

- Test för fixa parametrar
  - "Vanliga" hypotestest / konfidensintervall baserade på skattningarna och standradavvikleserna för skattningarna från IGLS (REML)
- Test för slumpmässiga parametrar (varianser och kovarianser)
  - Z-test (vid stora stickprov)
  - Bättre med Likelhood ratio test

$$D_{01} = -2LL(\lambda_0/\lambda_1)$$

där λ_0, λ_1 är likelihood under noll respektive alternativhypotesen. D_{01} är χ^2-fördelad med q frihetsgrader, där q är skillnaden i antalet skattade parametrar under de två modellerna.

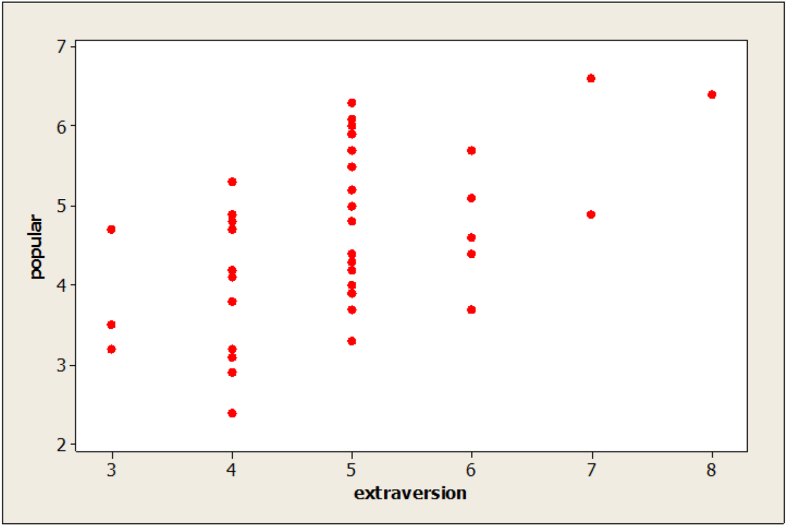
- Vid REML-skattning bör likelihood ratio-testet endast användas vid jämförande av modeller som skiljer sig åt endast genom slumpmässiga parametrar (ej fixa).

Exempel (data från Hox (2010). Multilevel Analysis)

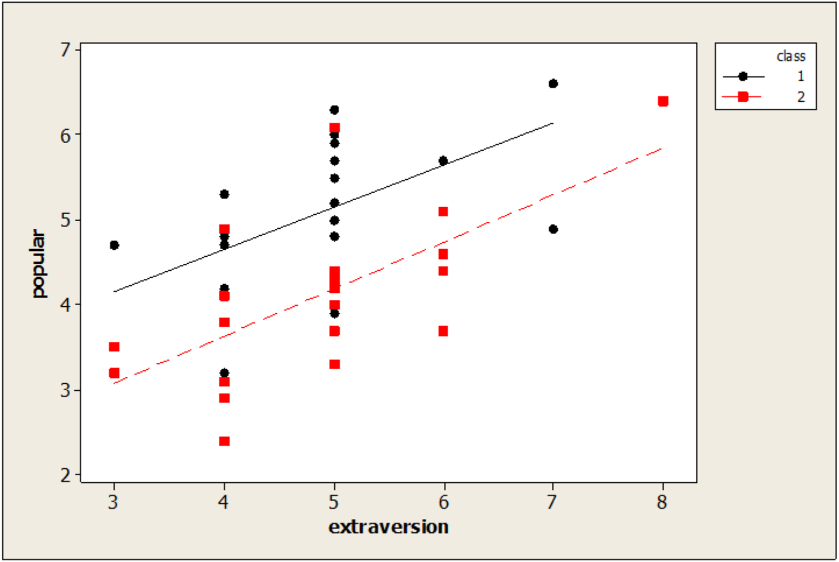
2000 elever från 100 olika klasser (skolor)

- Popularity (skala 0-10, elevnivå)
- Sex (elevnivå)
- Extraversion (skala 0-10, elevnivå)
- Teacher experience (År, klassnivå)

Spridningsdiagram mellan popularity-scores och extraversion för eleverna i de två första klasserna



Uppdelat på de två klasserna



Varianskomponent-modell

Basmodell

$$
\begin{aligned}
popular_{ij} &= \beta_{0j} + \epsilon_{ij}, \\
\beta_{0j} &= \beta_0 + u_{0j}
\end{aligned}
$$

$$popular_{ij} = \beta_0 + u_{0j} + \epsilon_{ij}$$

$$
\begin{aligned}
&\epsilon_{0ij} \sim N(0, \sigma_e^2) \\
&u_{0j} \sim N(0, \sigma_{u0}^2)
\end{aligned}
$$

Data

Obs	class	pupil	cons	extrav	sex	texp	popular	popteach	Zextrav	Zsex	Ztexp	Zpopular	Zpopteach	Cextrav	Ctexp	Csex
1	1	1	1	5	1	24	6.3	6	-0.17031	0.98881	1.48615	0.88501	0.66906	-0.215	9.737	0.4945
2	1	2	1	7	0	24	4.9	5	1.41401	-1.01081	1.48615	-0.12763	-0.04308	1.785	9.737	-0.5055
3	1	3	1	4	1	24	5.3	6	-0.96248	0.98881	1.48615	0.16170	0.66906	-1.215	9.737	0.4945
4	1	4	1	3	1	24	4.7	5	-1.75464	0.98881	1.48615	-0.27229	-0.04308	-2.215	9.737	0.4945
5	1	5	1	5	1	24	6.0	6	-0.17031	0.98881	1.48615	0.66802	0.66906	-0.215	9.737	0.4945
6	1	6	1	4	0	24	4.7	5	-0.96248	-1.01081	1.48615	-0.27229	-0.04308	-1.215	9.737	-0.5055
7	1	7	1	5	0	24	5.9	5	-0.17031	-1.01081	1.48615	0.59569	-0.04308	-0.215	9.737	-0.5055
8	1	8	1	4	0	24	4.2	5	-0.96248	-1.01081	1.48615	-0.63395	-0.04308	-1.215	9.737	-0.5055
9	1	9	1	5	0	24	5.2	5	-0.17031	-1.01081	1.48615	0.08937	-0.04308	-0.215	9.737	-0.5055
10	1	10	1	5	0	24	3.9	3	-0.17031	-1.01081	1.48615	-0.85095	-1.46737	-0.215	9.737	-0.5055
11	1	11	1	5	1	24	5.7	5	-0.17031	0.98881	1.48615	0.45102	-0.04308	-0.215	9.737	0.4945
12	1	12	1	5	1	24	4.8	5	-0.17031	0.98881	1.48615	-0.19996	-0.04308	-0.215	9.737	0.4945
13	1	13	1	5	0	24	5.0	5	-0.17031	-1.01081	1.48615	-0.05530	-0.04308	-0.215	9.737	-0.5055
14	1	14	1	5	1	24	5.5	6	-0.17031	0.98881	1.48615	0.30636	0.66906	-0.215	9.737	0.4945
15	1	15	1	5	1	24	6.0	5	-0.17031	0.98881	1.48615	0.66802	-0.04308	-0.215	9.737	0.4945
16	1	16	1	6	1	24	5.7	5	0.62185	0.98881	1.48615	0.45102	-0.04308	0.785	9.737	0.4945
17	1	17	1	4	0	24	3.2	2	-0.96248	-1.01081	1.48615	-1.35727	-2.17951	-1.215	9.737	-0.5055
18	1	18	1	4	0	24	3.1	3	-0.96248	-1.01081	1.48615	-1.42960	-1.46737	-1.215	9.737	-0.5055
19	1	19	1	7	1	24	6.6	7	1.41401	0.98881	1.48615	1.10201	1.38120	1.785	9.737	0.4945
20	1	20	1	4	0	24	4.8	4	-0.96248	-1.01081	1.48615	-0.19996	-0.75523	-1.215	9.737	-0.5055
21	2	21	1	8	1	14	6.4	6	2.20617	0.98881	-0.04014	0.95734	0.66906	2.785	-0.263	0.4945
22	2	22	1	4	0	14	2.4	3	-0.96248	-1.01081	-0.04014	-1.93592	-1.46737	-1.215	-0.263	-0.5055

Varianskomponent-modell

SAS

```
proc mixed data=popdata covtest;
class class;
 model popular= /solution;
 random intercept / subject=class;
 run;
```

The Mixed Procedure

Model Information

Data Set	WORK.POPDATA
Dependent Variable	popular
Covariance Structure	Variance Components
Subject Effect	class
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
class	100	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

Dimensions

Covariance Parameters	2
Columns in X	1
Columns in Z Per Subject	1
Subjects	100
Max Obs Per Subject	26

Number of Observations

Number of Observations Read	2000
Number of Observations Used	2000
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	6975.50576364	
1	2	6330.50966993	0.00000008
2	1	6330.50956329	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	class	0.7021	0.1086	6.46	<.0001
Residual		1.2218	0.03964	30.82	<.0001

Fit Statistics

-2 Res Log Likelihood	6330.5
AIC (smaller is better)	6334.5
AICC (smaller is better)	6334.5
BIC (smaller is better)	6339.7

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	645.00	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.0779	0.08739	99	58.10	<.0001

Resultat: skattning och test

$\widehat{\beta_0} = 5.0779(0.07839),\ p < .0001$

$\widehat{\sigma_e^2} = 1.2218(0.03964),\ p < .0001$
$\widehat{\sigma_{u0}^2} = 0.7021(0.1086),\ p < .0001$

Intraklass-korrelation: $\widehat{\rho} = \frac{.702}{.702+1.221} = 0.365$

Varianskomponent-modell med kön (sex) på individnivå - fix effekt

Vi lägger till kön (sex) som fix effekt

$$\begin{aligned} popularity_{ij} &= \beta_{0j} + \beta_1 sex_{ij} + \epsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \end{aligned}$$

$$popularity_{ij} = \beta_0 + u_{0j} + \beta_1 sex_{ij} + \epsilon_{ij}$$

$$\epsilon_{0ij} \sim N(0, \sigma_e^2)$$
$$u_{0j} \sim N(0, \sigma_{u0}^2)$$

Varianskomponent-modell med kön (sex) på individnivå - fix effekt

SAS

```
proc mixed data=popdata covtest;
 class class;
 model popular =  Csex / solution;
 random intercept  / subject=class;
 run;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	class	0.4838	0.07494	6.46	<.0001
Residual		0.8308	0.02696	30.81	<.0001

Fit Statistics

-2 Res Log Likelihood	5564.1
AIC (smaller is better)	5568.1
AICC (smaller is better)	5568.1
BIC (smaller is better)	5573.3

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	635.96	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.0769	0.07251	99	70.02	<.0001
Csex	1.3501	0.04403	1899	30.66	<.0001

Resultat: skattning och test

$$\widehat{\beta_0} = 5.0769(0.07251),\ p < .0001$$
$$\widehat{\beta_1} = 1.3501(0.04403),\ p < .0001$$

$$\widehat{\sigma_e^2} = 0.8308(0.02696),\ p < .0001$$
$$\widehat{\sigma_{u0}^2} = 0.4838(0.07494),\ p < .0001$$

Multilevel-modell med kön (sex) på individnivå - fix och slumpmässig effekt

Vi lägger till kön (sex) också som slumpmässig effekt

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

$$
popularity_{ij} = \beta_0 + u_{0j} + \beta_1 sex_{ij} + u_{1j} sex_{ij} + \epsilon_{ij}
$$

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

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

Multilevel-modell med kön (sex) på individnivå - fix och slumpmässig effekt

SAS

```
proc mixed data=popdata covtest;
 class class;
 model popular =  Csex / solution;
 random intercept  Csex/ subject=class type=un;
 run;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	class	0.4824	0.07554	6.39	<.0001
UN(2,1)	class	-0.04911	0.03999	-1.23	0.2194
UN(2,2)	class	0.05461	0.03552	1.54	0.0621

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Residual		0.8190	0.02720	30.11	<.0001

Fit Statistics

-2 Res Log Likelihood	5559.1
AIC (smaller is better)	5567.1
AICC (smaller is better)	5567.1
BIC (smaller is better)	5577.5

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
3	640.98	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.0803	0.07250	99	70.07	<.0001
Csex	1.3522	0.05030	87	26.88	<.0001

Resultat: skattning och test

$$\widehat{\beta_0} = 5.00803(0.07250),\ p < .0001$$
$$\widehat{\beta_1} = 1.3522(0.05030),\ p < .0001$$

$$\widehat{\sigma_e^2} = 0.8190(0.02720),\ p < .0001$$
$$\widehat{\sigma_{u0}^2} = 0.4824(0.07554),\ p < .0001$$
$$\widehat{\sigma_{u1}^2} = 0.05461(0.03552),\ p = .0621$$
$$\widehat{\sigma_{u01}} = -0.04911(0.03999),\ p = .2194$$

$$D_{01} = 5564.1 - 5559.1 = 5$$
$$q = 2$$

Varianskomponent-modell med individnivåvariabler - fixa effekter

Vi lägger till extraversion som fix effekt

$$
\begin{aligned}
popularity_{ij} &= \beta_{0j} + \beta_1 sex_{ij} + \beta_2 extra_{ij} + \epsilon_{ij} \\
\beta_{0j} &= \beta_0 + u_{0j}
\end{aligned}
$$

$$popularity_{ij} = \beta_0 + u_{0j} + \beta_1 sex_{ij} + \beta_2 extra_{ij} + \epsilon_{ij}$$

$$
\begin{aligned}
&\epsilon_{0ij} \sim N(0, \sigma_e^2) \\
&u_{0j} \sim N(0, \sigma_{u0}^2)
\end{aligned}
$$

Varianskomponent-modell med individnivåvariabler - fixa effekter

SAS

```
proc mixed data=popdata covtest;
 class class;
 model popular = Csex Cextrav / solution;
 random intercept / subject=class;
 run;
```

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
UN(1,1)	**class**	0.6271	0.09386	6.68	<.0001
Residual		0.5921	0.01922	30.80	<.0001

Fit Statistics	
-2 Res Log Likelihood	4948.3
AIC (smaller is better)	4952.3
AICC (smaller is better)	4952.3
BIC (smaller is better)	4957.5

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
1	1037.03	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	5.0773	0.08105	99	62.64	<.0001
Csex	1.2530	0.03743	1898	33.48	<.0001
Cextrav	0.4416	0.01616	1898	27.33	<.0001

Resultat: skattning och test

$$\widehat{\beta_0} = 5.0773(0.08105),\ p < .0001$$
$$\widehat{\beta_1} = 1.2530(0.08743),\ p < .0001$$
$$\widehat{\beta_2} = 0.4416(0.01616),\ p < .0001$$
$$\widehat{\sigma_e^2} = 0.5921(0.01922),\ p < .0001$$
$$\widehat{\sigma_{u0}^2} = 0.6271(0.09386),\ p < .0001$$

Multilevel-modell med individnivåvariabler: kön (fix) och extraversion (fix och slumpmässig)

Vi lägger till extraversion också som slumpmässig effekt

$$popularity_{ij} = \beta_{0j} + \beta_1 sex_{ij} + \beta_{2j} extra_{ij} + \epsilon_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{2j} = \beta_2 + u_{2j}$$

$$popularity_{ij} = \beta_0 + u_{0j} + \beta_1 sex_{ij} + \beta_2 extra_{ij} + u_{2j} extra_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$$
$$\begin{bmatrix} u_{0j} \\ u_{2j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u02} \\ \sigma_{u20} & \sigma_{u2}^2 \end{bmatrix}\right)$$

Multilevel-modell med individnivåvariabler: kön (fix) och extraversion (fix och slumpmässig)
SAS

```
proc mixed data=popdata covtest;
 class class;
 model popular = Csex Cextrav / solution;
 random intercept Cextrav/ subject=class type=un;
 run;
```

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	**class**	0.6744	0.1020	6.61	<.0001
UN(2,1)	**class**	-0.1052	0.02259	-4.65	<.0001
UN(2,2)	**class**	0.02979	0.007911	3.76	<.0001
Residual		0.5540	0.01853	29.89	<.0001

Fit Statistics	
-2 Res Log Likelihood	4873.1
AIC (smaller is better)	4881.1
AICC (smaller is better)	4881.1
BIC (smaller is better)	4891.5

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
3	1112.28	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	5.0238	0.08413	99	59.71	<.0001
Csex	1.2444	0.03651	1799	34.08	<.0001
Cextrav	0.4438	0.02344	99	18.93	<.0001

Resultat: skattning och test

$$\widehat{\beta_0} = 5.0238(0.08413),\ p < .0001$$
$$\widehat{\beta_1} = 1.2444(0.03651),\ p < .0001$$
$$\widehat{\beta_2} = 0.4438(0.02344),\ p < .0001$$

$$\widehat{\sigma_e^2} = 0.5540(0.01853),\ p < .0001$$
$$\widehat{\sigma_{u0}^2} = 0.6744(0.1020),\ p < .0001$$
$$\widehat{\sigma_{u2}^2} = 0.02979(0.007911),\ p < .0001$$
$$\widehat{\sigma_{u02}} = -0.1052(0.02259),\ p < .0001$$

$$D_{01} = 4948.3 - 4873.1 = 75.2$$
$$q = 2$$

Multilevel-modell med individnivåvariabler kön (fix), extraversion (fix, random) och teacher experience (klassnivå)

Vi lägger till teacher experience på klassnivå

$$
\begin{aligned}
popularity_{ij} &= \beta_{0j} + \beta_1 sex_{ij} + \beta_{2j} extra_{ij} + \epsilon_{ij} \\
\beta_{0j} &= \beta_0 + \beta_3 tex_j + u_{0j} \\
\beta_{2j} &= \beta_2 + \beta_4 tex_j + u_{2j}
\end{aligned}
$$

$$popularity_{ij} =$$
$$\beta_0 + \beta_3 tex_j + u_{0j} + \beta_1 sex_{ij} + \beta_2 extra_{ij} + \beta_4 tex_j extra_{ij} + u_{2j} extra_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N\left(0, \sigma_e^2\right)$$
$$\begin{bmatrix} u_{0j} \\ u_{2j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u02} \\ \sigma_{u20} & \sigma_{u2}^2 \end{bmatrix} \right)$$

Multilevel-modell med individnivåvariabler kön (fix), extraversion (fix, random) och teacher experience (klassnivå)

SAS

```
proc mixed data=popdata covtest;
 class class;
 model popular =  Csex Ctexp Cextrav Ctexp*Cextrav/ solution;
 random intercept Cextrav / subject=class type=un;
 run;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	class	0.2859	0.04526	6.32	<.0001
UN(2,1)	class	-0.00437	0.009726	-0.45	0.6531
UN(2,2)	class	0.005409	0.004411	1.23	0.1101
Residual		0.5528	0.01842	30.00	<.0001

Fit Statistics

-2 Res Log Likelihood	4780.5
AIC (smaller is better)	4788.5
AICC (smaller is better)	4788.5
BIC (smaller is better)	4798.9

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
3	570.55	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	4.9952	0.05659	98	88.27	<.0001
Csex	1.2407	0.03623	1799	34.24	<.0001
Ctexp	0.09724	0.008690	1799	11.19	<.0001
Cextrav	0.4509	0.01742	98	25.89	<.0001
Ctexp*Cextrav	-0.02473	0.002555	1799	-9.68	<.0001

Resultat: skattning och test av parametrar

$$\widehat{\beta_0} = 4.9952(0.05659),\ p < .0001$$
$$\widehat{\beta_1} = 1.2407(0.03623),\ p < .0001$$
$$\widehat{\beta_2} = 0.4509(0.01742),\ p < .0001$$
$$\widehat{\beta_3} = 0.09724(0.008690),\ p < .0001$$
$$\widehat{\beta_4} = -0.02473(0.002555),\ p < .0001$$

$$\widehat{\sigma_e^2} = 0.5528(0.01842),\ p < .0001$$
$$\widehat{\sigma_{u0}^2} = 0.2859(0.04526),\ p < .0001$$
$$\widehat{\sigma_{u2}^2} = 0.005409(0.004411),\ p = .1101$$
$$\widehat{\sigma_{u02}} = -0.00437(0.009726),\ p = .6531$$

Fler än två nivåer: föregående modeller kan enkelt utvecklas till fler nivåer

$$
\begin{aligned}
Y_{ijh} &= \beta_{0jh} + \beta_{1jh}x_{ijh} + \epsilon_{ijh} \\
\beta_{0jh} &= \beta_{0h} + u_{0jh} \\
\beta_{1jh} &= \beta_{1h} + u_{1jh} \\
\beta_{0h} &= \beta_0 + v_{0h} \\
\beta_{1h} &= \beta_1 + v_{1h}
\end{aligned}
$$

ijh är tex i:te eleven i j:te klassen i h:te området.

$$y_{ijh} = \beta_0 + \beta_1 x_{ijh} + v_{0h} + v_{1h}x_{ijh} + u_{0jh} + u_{1jh}x_{ijh} + \epsilon_{ijh}$$

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

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

$$\begin{bmatrix} v_{0h} \\ v_{1h} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01} \\ \sigma_{v10} & \sigma_{v1}^2 \end{bmatrix}\right)$$

Exempel på SAS-kod 3 nivåer, varianskomponentmodell

```
proc mixed data=popdata2 covtest;
 class class area;
 model popular =   / solution;
 random intercept  / subject=area;
 random intercept  / subject=class(area);
 run;
```