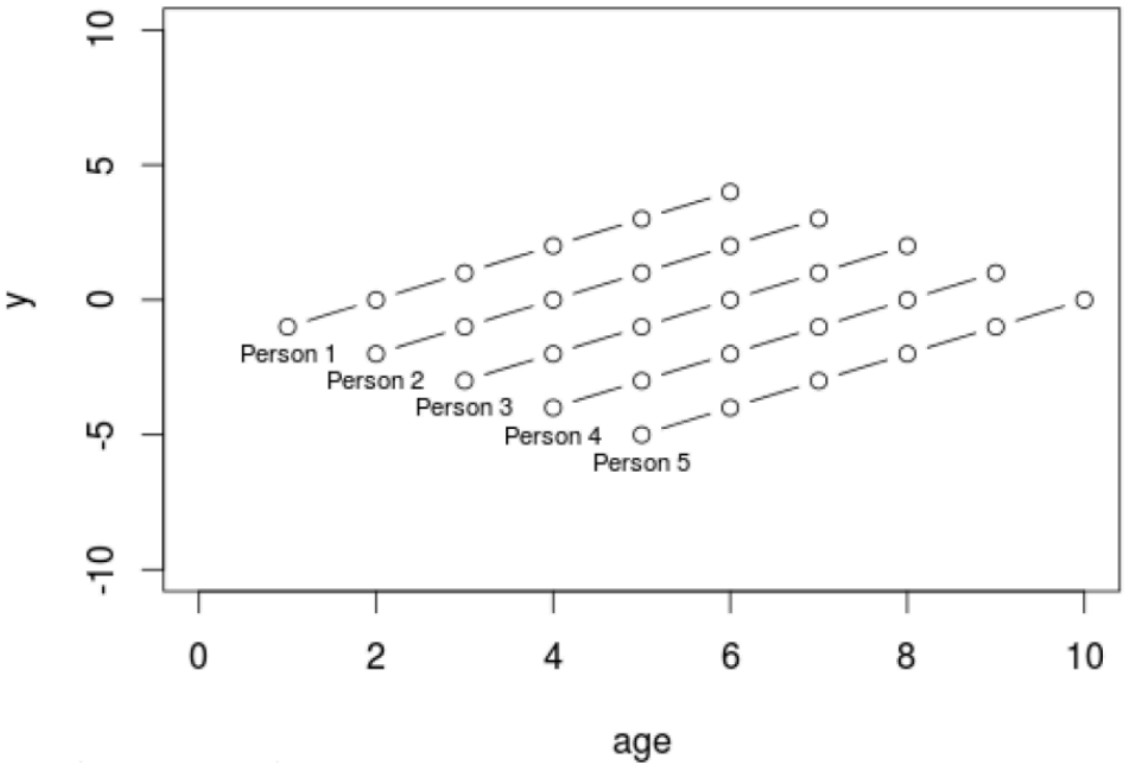
GLMM: Generalised Linear Mixed Models för longitudinell data

> För denna föreläsning rekommenderas följande litteratur:
> - Bok: Goldstein, kapitel 6.
> - Artikel: Singer 1998 (andra halvan).
> - Artikel: Tan, Kan, Hogan 2010.

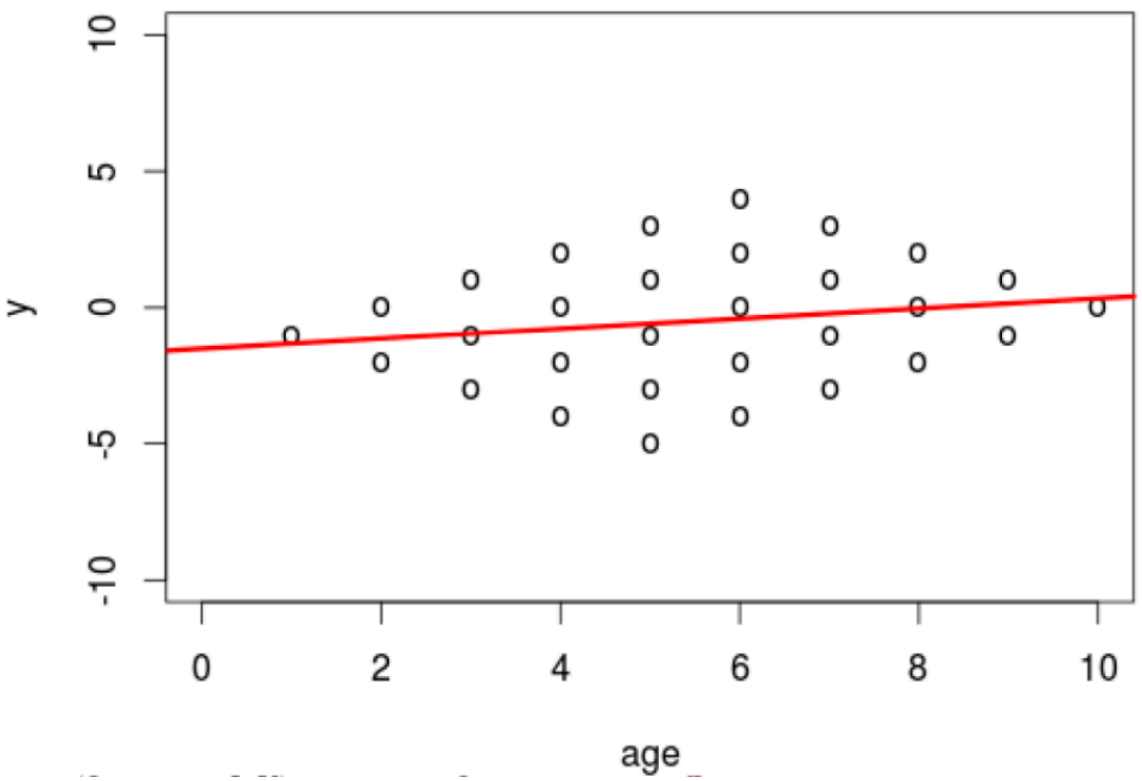
Introduktion till longitudinell data

Vad är longitudinell data?

Longitudinell data är data som samlas in genom att upprepade gånger mäta samma urval, som inidivider eller hushål under en längre tidsperiod för att analysera förändringar och mönster över tid. Detta skiljer sig från tvärsnittsdata, där data samlas in från olika urval vid en specifik tidpunkt. På grund av att longitudinell data är mätt på samma urval krävs speciella metoder för att kunna spåra utveckling och trender, vilket kan vara resurskrävande och känsligt för bortfall — de som slutar delta i studien.



Ovan ses ett diagram över ett specifikt urval av fem individer som mätts under en tidsperiod. Om man endast skulle analysera datamaterialet utan att **ta hänsyn till individer** skulle man få en falsk uppfattning av verkligheten. Detta skulle t.ex se ut som:



Longitudinell data - analys och presentation av datamaterial

I figur 1 nedan visas upprepade mätningar av tandutvecklingsdata från *Pathoff & Roy (1964)*. Det är alltså ingen bild på av personer, utan en graf över mätvärden.

På x-axeln ser man ålder från 8-14 år. På y-axeln ses den genomsnittliga längden på en specifik kindtand, mätt i militer. Figur 1 visar flickor i ålder 1-11 (eller snarare observationer märkta som “girls”) som gröna linjer. De blå linjerna visar pojkar från 12-27 (även här är det ettiketer, inte faktiskt åldrar — det är individnummer). Varje linje representerar en individ, där inidivden mätts vid flera tidpunkter. Därför syns flera gröna linjer: en per flick/pojke. Man kan notera en ökning i tandlängd över tid; positiv utvekling.

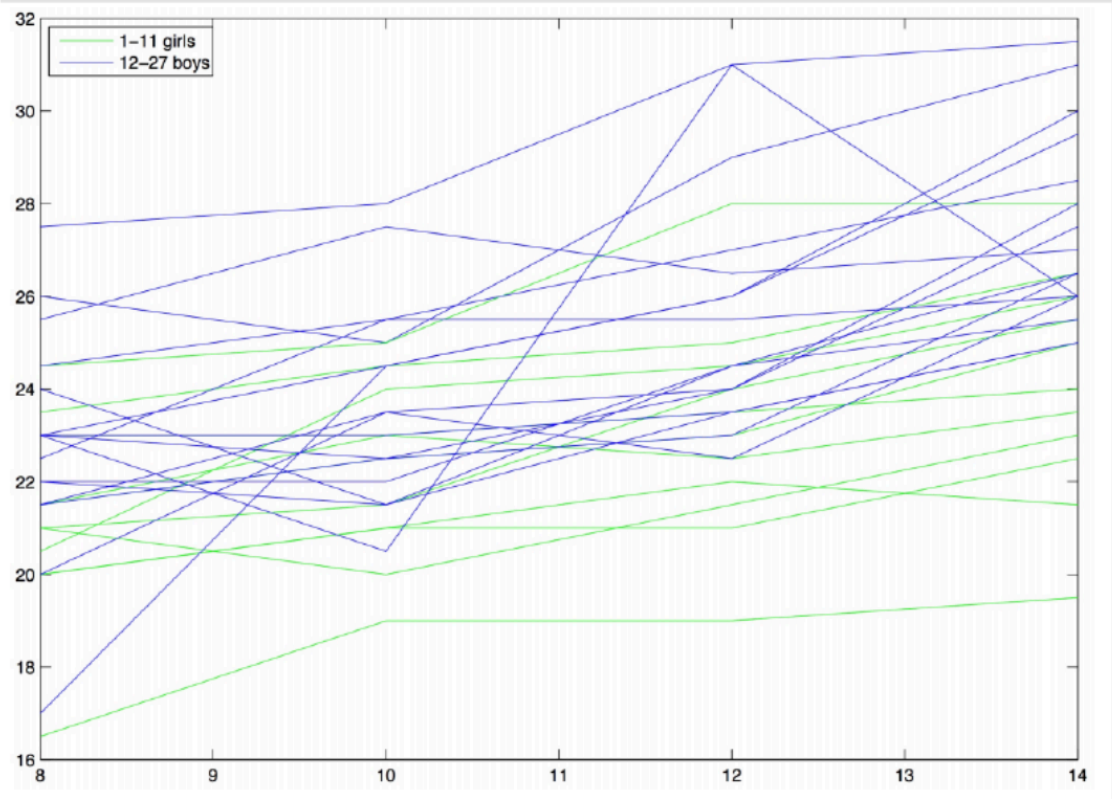


Figure 1: Dental data from Pothoff and Roy (1964)

Presentation i tabell:

I en tabell visas datamaterialet där man kan se (t.ex) $Y_{1,1} = 21$ vid första tidpunkten. Detta innebär individ 1 vid tidpunkt 1. $Y_{1,4} = 23$. Detta är alltså individ 1 vid tidpunkt 4.

Table 3.1: *Dental data*

id	gender	t_1	t_2	t_3	t_4	id	gender	t_1	t_2	t_3	t_4
1	F	21.0	20.0	21.5	23.0	12	M	26.0	25.0	29.0	31.0
2	F	21.0	21.5	24.0	25.5	13	M	21.5	22.5	23.0	26.0
3	F	20.5	24.0	24.5	26.0	14	M	23.0	22.5	24.0	27.0
4	F	23.5	24.5	25.0	26.5	15	M	25.5	27.5	26.5	27.0
5	F	21.5	23.0	22.5	23.5	16	M	20.0	23.5	22.5	26.0
6	F	20.0	21.0	21.0	22.5	17	M	24.5	25.5	27.0	28.5
7	F	21.5	22.5	23.0	25.0	18	M	22.0	22.0	24.5	26.5
8	F	23.0	23.0	23.5	24.0	19	M	24.0	21.5	24.5	25.5
9	F	20.0	21.0	22.0	21.5	20	M	23.0	20.5	31.0	26.0
10	F	16.5	19.0	19.0	19.5	21	M	27.5	28.0	31.0	31.5
11	F	24.5	25.0	28.0	28.0	22	M	23.0	23.0	23.5	25.0
						23	M	21.5	23.5	24.0	28.0
						24	M	17.0	24.5	26.0	29.5
						25	M	22.5	25.5	25.5	26.0
						26	M	23.0	24.5	26.0	30.0
						27	M	22.0	21.5	23.5	25.0

Figure 2: Dental data from Pothoff and Roy (1964)

Upprepade observationer över tid

Eftersom det är upprepade observationer på samma individer över tid ⇒ man har samma person (eller företag, klass, land etc) mätta flera gånger. Det innebär att man har en tidsvariabel t som anger tidpunkten observationen var mätt. Konsekvent innebär detta att datan är ordnad i tid — inte bara slumpmässiga mätningar. Man kan uttrycka detta som

> Beroende mätningar vid olika tidpunkter ⇒ autokorrelation

Mätningar från samma individ vid olika tidpunkter är inte oberoende. Tänk, om en person har hög stress vid tidpunkt 1, är det ganska sannolikt att personen har relativt hög stress vid tidpunkt 2 också (jämfört med andra individer). **Autokorrelation** innebär att det finns korrelation mellan värden av samma variabel vid olika tidpunkter, (t.ex) att korrelationen mellan stress vid månad 1 och stress vid månad 2 för samma person. **Detta bryter mot antagandet i vanligt linjär regression/ANOVA att alla observationer är oberoende.** Konsekvent innebär detta att dessa typer av metoder inte går att använda.

Vad händer om man ändå provar regression eller ANOVA för longitudinell data?

Om man ignorerar att observationerna är beroende (autokorrelarade) och kör en vanlig regression som om alla rader i datan är oberoende kommer:

- Standradfel bli fel.
- P-värden blir skeva.
- Man drar fel slutsatser

På grund av detta vill man använda andra typer av metoder såsom:

- Mixed models / Multilevel modeller.

Jämförelse mellan hierarkiska metoder?

På grund av den longitudinella datan; t.ex om man mäter 5 personers IQ mellan ett tidsintervall kan det komma att finnas **mellan individ** effekt och **inom individ** effekt. Vilket innebär att sättet att analysera longitudinell data är mycket lik det sätt man analyserar hierarkisk data.

Repetion av hierarkisk data

I hierarkisk data mäter man enheter grupperade i olika nivåer (bl.a):

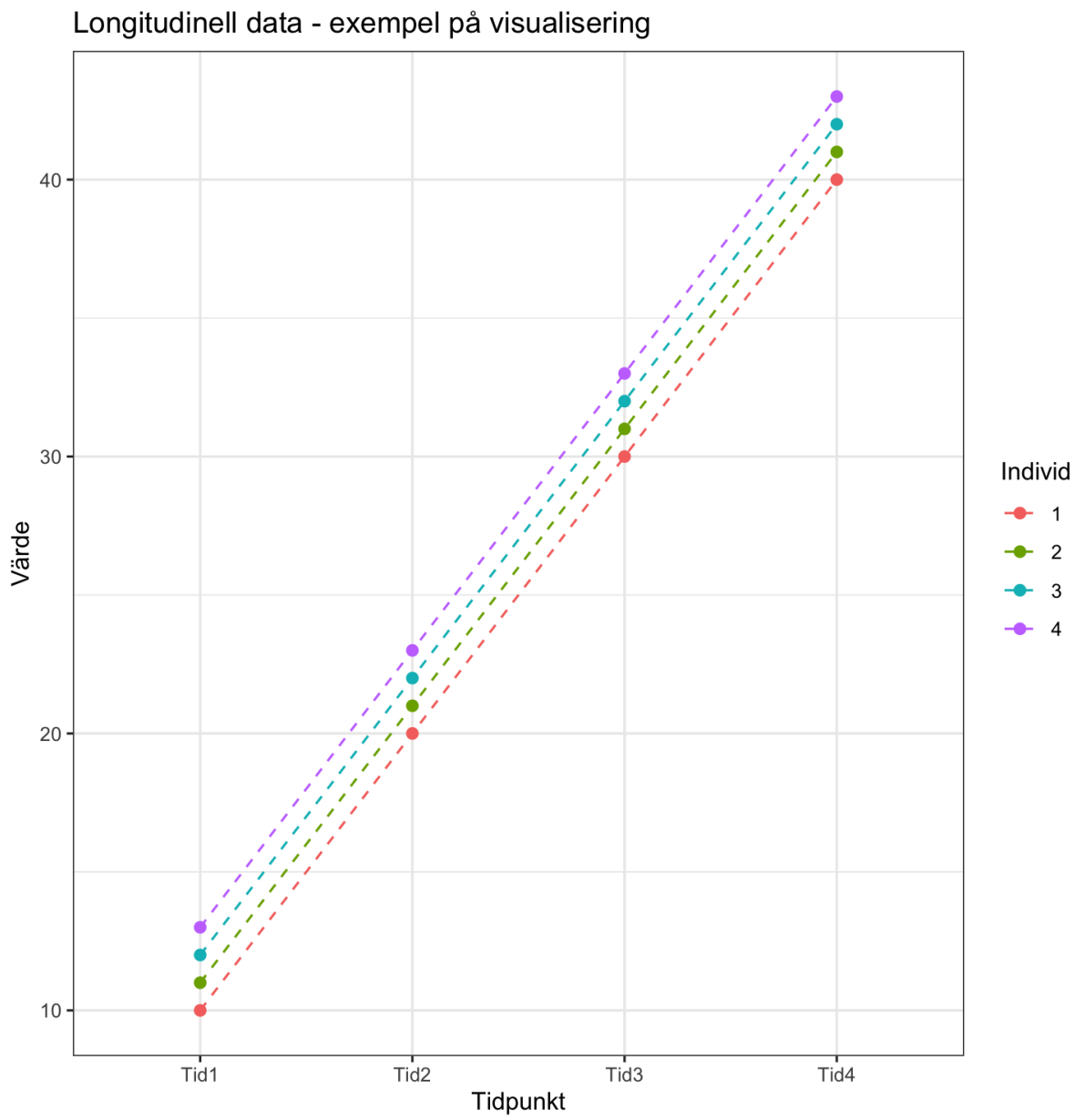
- Nivå 1: Elever.
- Nivå 2: Klasser.
- Nivå 3: Skolor.

Observationerna inom en grupp tenderar att likna varandra. Ignorerar man dessa grupperingar kan man vissa **mycket viktiga gruppeffekter.** Ignorerar man korrelationer mellan observationer inm grupper får an missvisande variansskattningar.

Longitudinell data

När man nu har upprepade observation på samma individ skapas en 2-nivå hierarki, där:

- Nivå 1: upprepade mätningar per individ (tid).
- Nivå 2: Individer.



Det är också möjligt att ha upprepningar i högre nivåer av data (t.ex årliga provresultat för elever i ett urval av skolor):

- Nivå 1: Elever
- Nivå 2: Upprepade mätningar per skola (tid)
- Nivå 3: Skola.

Hierarkisk data - modell

I hierarkisk form skulle modellen formuleras enligt:

$$\text{Nivå 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}$$
$$\text{Nivå 2: } \beta_{0j} = \beta_0 + u_{oj} \quad \beta_{1j} = \beta_1 + u_{1j}$$

Vi har alltså ett slumpmässigt intercept och slumpmässig lutning för varje individ. Detta är rätt så självklart i diagrammet ovan där data visualiserats. Man ser att lutningen i princip är densamma för alla linjer (eftersom de är parallella), men interceptet är olika för vardera individ. Varför det är såhär kan inte förklaras för denna datan.

I kombinerad form skulle denna ekvation kunna skrivas direkt som:

$$Y_{ij} = (\beta_0 + \beta_1 X_{ij}) + (u_{0j} + u_{1j}X_{ij} + \epsilon_{ij})$$

Med följande antaganden:

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

I denna modell skattas alltså två fixa effekter och fyra slumpmässiga parametrar. Detta innebär att när man kollar på ekvationsformuleringen är β_0, β_1 den fasta interceptet (genomsnittligt intercept i i populationen) och fasta lutningen (genomsnittlig effekt av X). Där

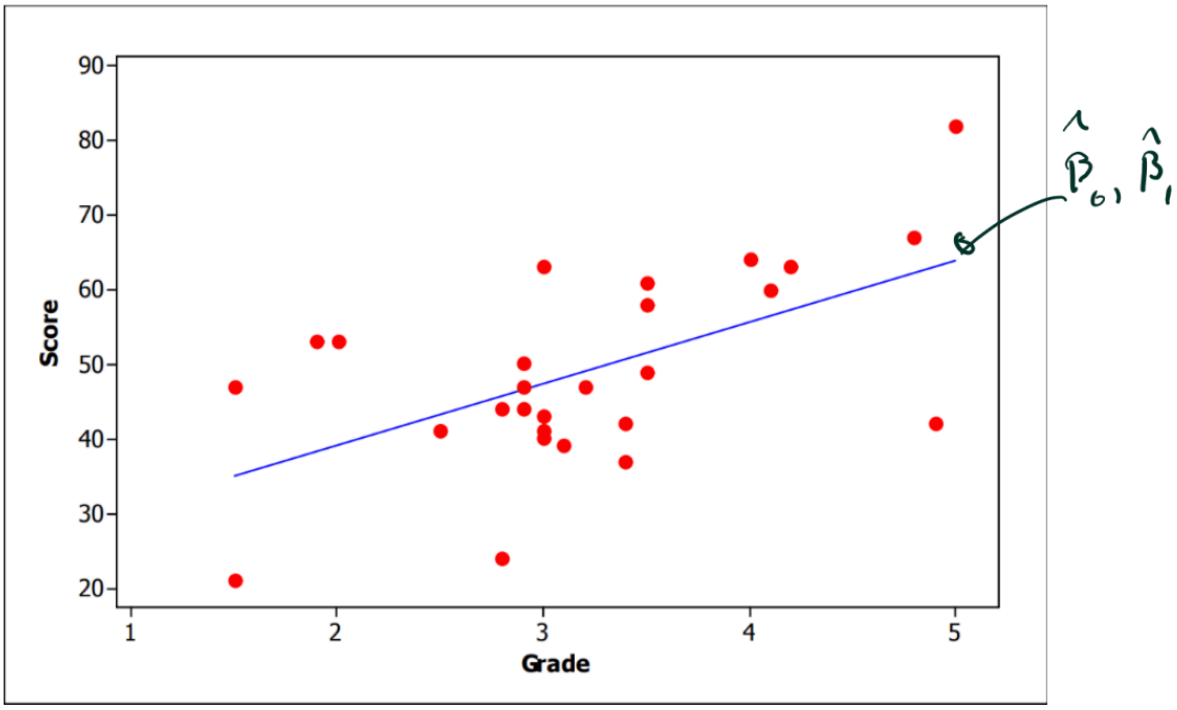
$$\text{Intercept: } \beta_{0j}$$
$$\text{Lutningen: } \beta_{1j}$$

Detta innebär i sin tur att

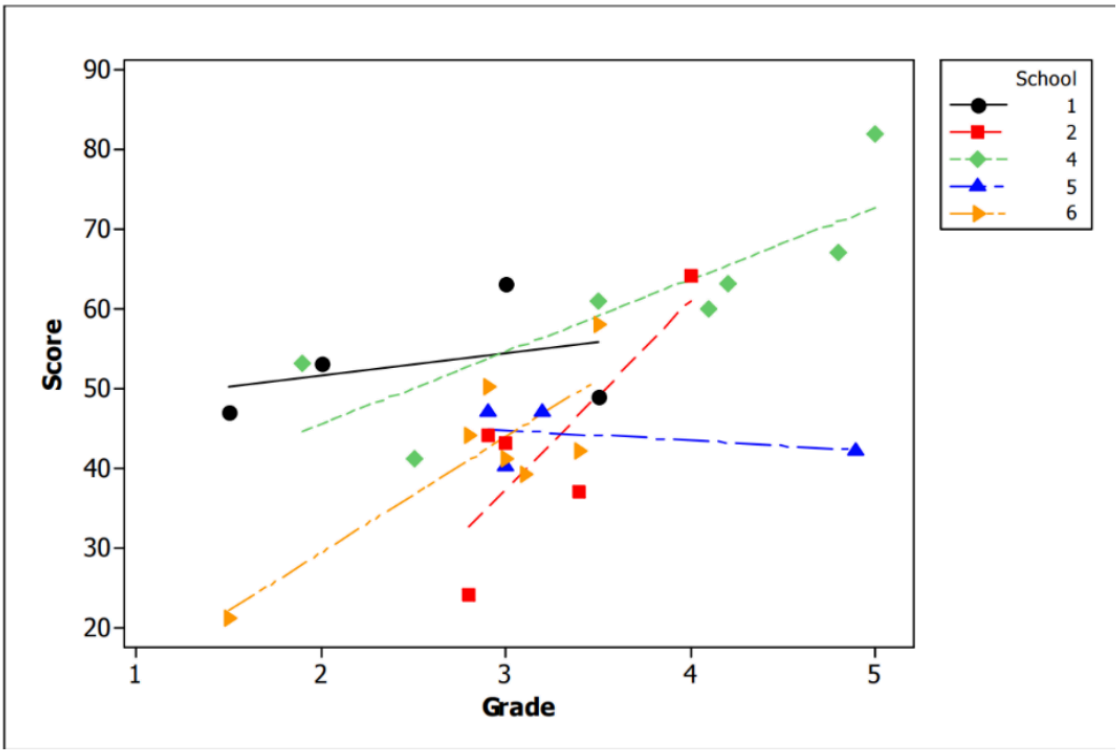
- β_0 är det fixa interceptet
- β_1 är den fixa lutningen
- u_{0j} är hur mycket interceptet för j avviker från β_0
- u_{1j} är hur mycket lutningen för j avviker från β_1

ALLTSÅ u är inte själva lutningen utan avvikelsen i lutning (eller intercept) för just den individen eller gruppen.

Skattningen för modellen ovan



Här ser man att elverna följer samma samband!



Men i denna bild är det annorlunda för nu ser man att elever inom samma skola har ett liknande samband men sambandet skiljer sig mellan skolor!

Longitudinell data - egenskaper

Autokorrelation:

- **Positiv autokorrelation:** detta innebär att om en person har ett högt värde vid tidpunkt t är det ganska troligt att hen också har ett relativt högt värde vid $t + 1$. Det vill säga värdena liknar sig själva över tid.
- **Minskar med tidsavstånd mellan två observationer:** alltså ju längre tid det går mellan två mätningar desto svagare blir sambandet mellan dem! Detta är rätt så självklart, man liknar sig själv mer om en månad än 10 år t.ex.
- **Korrelation mellan mycket långa tidsavstånd är ofta skild från noll:** innebär även om urvalet glider isär över långt tid så är människor ofta ändå lite lika sig själv: t.ex kroppslängd, blodtryck, skolreslutat ⇒ det finns rest av korrelation kvar.

- **Korrelation mellan mycket korta tidsavstånd är sällan nära ett:** även om två mätningar ligger nära i tid är de inte identiska. Det finns alltid (nästan) mätfel, dagsform, slump. Därför är ofta korrelationen hög men aldrig 1.

Saknad data:

- **Saknade mättillfällen:** en individ missar enstaka times t.ex glömmer fylla i enkäten ett år eller dyker inte upp på ett besök, men är där senare.
- **Drop-outs:** personen slutuat helt i studien. Flyttar, vill ınte delta längre, dör osv. Detta är extra jobbigt — ofta är det inte slumpmässigt.
- **Överlevare:** efter ett tag är det bara en viss grupp kvar (t.ex), de som orkar fortsätta, de som inte dött/flyttat, de som är mest motiverade. I detta fall bygger analysen på en "särskilt" del av ursprungsgruppen ⇒ **survivor bias/ överlevar bias.**

Longitudinell data: modeller (viktigaste delen)

Nu börjar föreläsningen på riktigt! Denna föreläsning fokuserar på **GLMM — Generaliserade Linjära Mixed Models,** där man kan ha fler än två nivåer. Nästa föreläsning fokuserar istället på faktoranalys approach. Denna approach är dock sämre vid olika antal tidpunkter, men bra när responsvariabeln är latent.

Generalised Linear Mixed Models

Nedan presenteras GLMM modeller, vilket är mycket bra eftersom man kan även modellera icke linjära samband med bland annat polynom. Skillnaden från denna typ av modell och modellen som används i hierarkisk data är att man nu har en matris **R** vilket är residualernas kovariansmatris ⇒ **hur felen är korrelerade över tid inom samma individ.** Den beskriv på detta sätt på grund av att samma individ mäts flera gånger så uppstår autokorrelation över tid.

Longitudinella data - GLMM

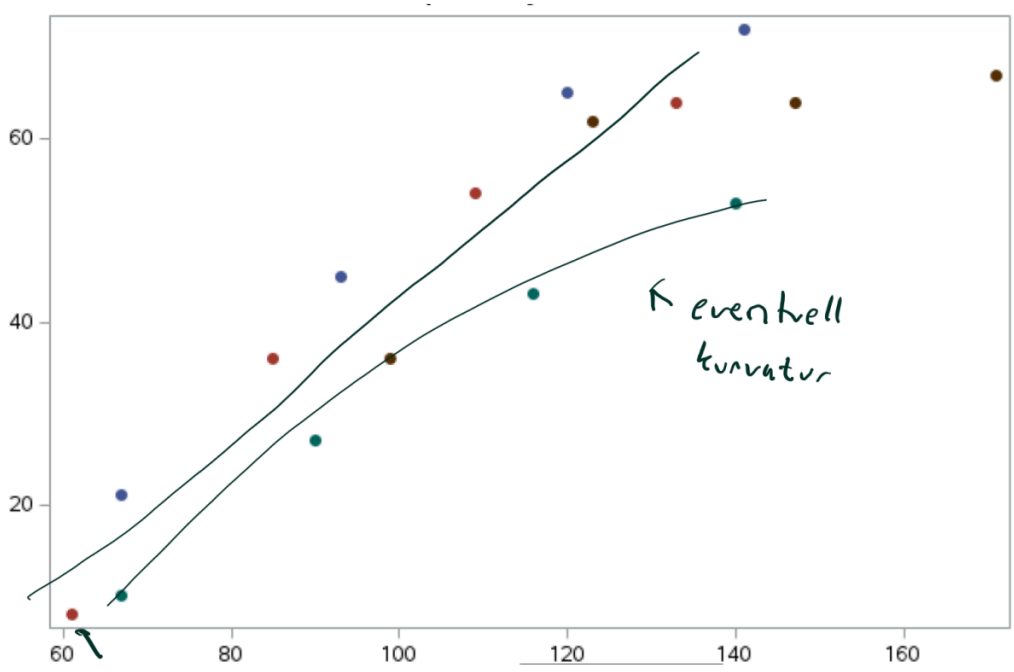
Enkel två-nivåmodell med tid som oberoende variabel

- Hierarkisk form:
  - Nivå 1: $Y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \varepsilon_{ij}$
  - Nivå 2: $\beta_{0j} = \beta_0 + u_{0j}$
    $\beta_{1j} = \beta_1 + u_{1j}$
- Kombinerad form:
  - $Y_{ij} = (\beta_0 + \beta_1 t_{ij}) + (u_{0j} + u_{1j} t_{ij} + \varepsilon_{ij})$
- $\boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \boldsymbol{R})$
- $\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix} \right)$

Två fixa, tre ”slumpmässiga” parametrar och ett antal parametrar att skatta i $\boldsymbol{R}-$matrisen beroende på dess struktur.

Exempel på hur den skattade linjen/linjerna ser ut är:

Longitudinella data - plot



Parameterestimering

I GLMM estimeras två typer av parametrar: **fixa effekter** och **varians-kovarians komponenter.** Man kanske hade kunnat tänka sig att ANOVA hade funkat i detta fallet, vilket ibland är ok men oftast i longitudinell ska det undvikas på grund av:

- Saknade värden: en individ försvinner.
- Tid kanske inte kan betraktas som en kategorisk variabel.
- Olika antal mätningar på individ ⇒ ej balanserad.

Därför är multilevel mixed models mer lämpliga!

- Full ML ger bias i variansskattningar.
- REML kan användas ist! gör väntevärdesriktiga skattningar!

Modellformulering av GLMM

Läs av sliden: tillräckligt bra förklarat:

Longitudinella data - GLMM

- **Matrisform:**
  $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$
  $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R})$
  $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G})$
  $Cov(\boldsymbol{u}, \boldsymbol{e}) = \boldsymbol{0}$
  Modellens parametrar består av fixa effekter i vektorn $\boldsymbol{\beta}$ och parametrar i matriserna $\boldsymbol{G}$ och $\boldsymbol{R}$.

- **Enkel två-nivåmodell med tid som oberoende variabel:**
  $$\begin{pmatrix} Y_{ij} \\ \vdots \\ Y_{N_ij} \end{pmatrix} = \begin{pmatrix} 1 & t_{ij} \\ \vdots & \vdots \\ 1 & t_{N_ij} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 & t_{ij} \\ \vdots & \vdots \\ 1 & t_{N_ij} \end{pmatrix} \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} + \begin{pmatrix} \varepsilon_{ij} \\ \vdots \\ \varepsilon_{N_ij} \end{pmatrix}$$
  där $N_i =$ antal upprepade mätningar för individ j
  $\beta_{0j} = \beta_0 + u_{0j}$
  $\beta_{1j} = \beta_1 + u_{1j}$

Strukturer för kovariansmatrisen R

R: UN (Unstructured covariance structure)

Nästa slide visar **R** vilket är **kovariansmatrisen för residualerna** inom en individ; alltså hur felen vid olika tidpunkter hänger ihop. Här är σ_i^2 variansen vid tidpunkt i och σ_{ki} är kovariansen mellan tidpunkt k och i. Följande teori behöver tas upp:

- UN: Unstructured innebär att alla elementen i matrisen får vara egna, fria parametrar. INgen förenkling som "samma korrealtion överallt" eller AR(1) osv. Det vill säga för tidpunkter blir antalet parametrar i R:

$$\frac{i(i+1)}{2}$$

om $i = 3$ fås 6 parametrar, $i = 5$ ger 15 parametrar, $i = 8$ ger 36 parametrar. **Det växer alltså snabbt.**

- Eftersom varje par av tidpunkter får sin egen kovarians (och därmed korrelation), blir det massor av parametrar att skatta när i är lite större. Det kräver MYCKET DATA och BRA SPRIDNING I TIDERNA. Annars blir skattningarna instabila

Strukturer för kovariansmatrisen $\boldsymbol{R}$ - Unstructured

$$\boldsymbol{R} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots \sigma_{1i} \\ \sigma_{12} & \sigma_2^2 & \dots \sigma_{2i} \\ \vdots & \vdots & \vdots \\ \sigma_{1i} & \sigma_{2i} & \dots \sigma_i^2 \end{pmatrix}$$

- Unstructured covariance structure (UN) är den mest komplicerade
- Unika korrelationer ger många parametrar att skatta
- SAS returnerar ofta ett felmeddelande för att det blir för många parametrar att skatta

R: CS (Compound Symmetry structure)

CS är den enklaste strukturen eftersom att alla mättillfällen har **samma varians.**

- Alla par av tidpunkter har samma kovarians/korrelation ρ oavsett hur långt det är mellan dem i tid.
- På grund av detta räcker det med två parametrar: en varians och en korrealtion.

Den högra skrivningen visar samma sak uppdelat i:

- en **mellan individ varians** (σ^2 t.ex slumpintercept).
- Plus en **residualvarians** (σ_e^2) på diagonalen.

Strukturer för kovariansmatrisen $\boldsymbol{R}$ - Compound Symmetry

$$\boldsymbol{R} = \sigma^2 \begin{pmatrix} 1 & \rho & \dots \rho \\ \rho & 1 & \dots \rho \\ \vdots & \vdots & \vdots \\ \rho & \rho & \dots 1 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \sigma_e^2 & \sigma_1 & \dots \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_e^2 & \dots \sigma_1 \\ \vdots & \vdots & \vdots \\ \sigma_1 & \sigma_1 & \dots \sigma^2 + \sigma_e^2 \end{pmatrix}$$

- Compound Symmetry (CS) structure är den enklaste
- Korrelerade feltermer mellan tidpunkter inom individer
- Korrelationerna är samma oavsett skillnaden i tid mellan de upprepade mätningarna

R: AR(1) (First Order Autoregressive)

En first order autoregressive innebär att man gör en tidsförskjutning. Rent teoretisk innebär detta att residualerna över tid är korrelerade, men korrelationen minskar ju längre ifrån varandra som tidpunkterna ligger. Detta innebär:

1. **Hög korrelation mellan mätningar nära varandra:** Tid t och tid $t + 1$ → korrelationen = ρ är hög
2. **Lägre korrelation ju längre apart de är:** tid t och tid $t + 2$ är korrealtionen = ρ^2 och för t och $t + 3$ → korrelationen = ρ^3.

Alltså:

- Nära i tid → hög korrelation.
- Långt isär i tid → låg korrelation.
- Aldrig noll så länge $\rho \neq 0$.

3. **Endast jämna tidsintervall:** AR(1) kräver att tiden är lika långt mellan varje mättillfälle. T.ex varje 6:e månad, varje år, varje vecka. Annars går inte hoppet ρ, ρ^2, ρ^3 logiskt ihop.
4. **Parametrar som skattas i AR(1):**
   - σ^2 (residualvariansen)
   - ρ (autokorrelationen)
   - → Mycket enklare är unstructured!

Sammanfattningsvis antar AR(1) att residualerna är starkt korrelerade mellan närliggande punkter och att korrelationen avtar exponetiellt med tidsavståndet.

Strukturer för kovariansmatrisen $\boldsymbol{R}$ -First Order Autoregressive AR(1)

$$\boldsymbol{R} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots \rho^{n_j-1} \\ \rho & 1 & \rho & \dots \rho^{n_j-2} \\ \vdots & \vdots & \vdots & \\ \rho^{n_j-1} & \rho^{n_j-2} & \rho^{n_j-3} & \dots 1 \end{pmatrix}$$

- Korrelationerna är som störst för intilligande tidpunkter
- Korrelationerna minskar systematiskt med ökad skillnad mellan tidpunkter
- Endast tillämplig på jämnt fördelade tidsintervall så att på varandra följande korrelationer är lika med ρ upphöjt till 1,2,3,4 osv.
- Korrelation mellan tid 1 och 2 blir $\rho^{t_2-t_1} = \rho$, mellan tid 1 och 3 $\rho^{t_3-t_1} = \rho^2$ osv

$\rho^{2-1} = \rho$

R: SP(POW) (Spatial Power)

Kort förklarat är Spatial Power en flexibel version av AR(1) som fungerar när mättidpunkterna inte ligger med jämna mellanrum. **Varför behövs denna om AR(1) finns?**

I AR(1) antas det att tiden är jämn t.ex år 1,2,3,4 ⇒ alltid 1 år mellan observationer. Men i longitudinella studier mäts individer vid:

- Olika tidpunkter.
- Olika dagar.
- Oregelbundna uppföljningsintervall.

Där av fungerar inte en AR(1).

Spatial Power fungerar genom att korrelationen mellan två tidpunkter t_k och t_l är $\rho^{|t_k - t_l|}$ alltså att korrelationen beror direkt på tidsavståndet. Men om tiderna är jämnd fördelade innebär det att en SP = AR(1). **Alltså SP är en generaliserad version av AR(1).**

Sammanfattningsvis:

- SP tillåter ojämna tidsintervall.
- Korrelationen mellan residualerna ges av $\rho^{\text{tidsavstånd}}$.
- SP = AR(1) om tidsavstånden är lika stora.
- Det är ofta det bästa valet i praktiska longitudinella studier.

Strukturer för kovariansmatrisen $\boldsymbol{R}$ - Spatial Power SP(POW)

$$\boldsymbol{R} = \sigma^2 \begin{pmatrix} 1 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} & \dots \rho^{|t_1-t_{n_j}|} \\ \rho^{|t_2-t_1|} & 1 & \rho^{|t_2-t_3|} & \dots \rho^{|t_2-t_{n_j}|} \\ \vdots & \vdots & \vdots & \\ \rho^{|t_{n_j}-t_1|} & \rho^{|t_{n_j}-t_2|} & \rho^{|t_{n_j}-t_3|} & \dots 1 \end{pmatrix}$$

- När tidpunkterna inte är jämnt fördelade så är spatial power ett ekvivalent alternativ till AR(1)
- Korrelationerna är upphöjda till de faktiska tidskillnaderna
- Om tidpunkterna är jämnt fördelade så är spatial power ekvivalent med AR(1)

R: Toeplitz

I Toeplitz har:

- Alla tidpunkter samma varians σ^2 på diagonalen.
- Korrelationen beror på avståndet i tid, men varje tidsavstånd har sin egen parameter.
  - Korrelation mellan närliggande tider = ρ_1
  - Mellan två steg isär = ρ_2
  - Mellan tre steg isär = ρ_3

Alltså är Toeplitz som AR(1) beroende av tidsavstånd, men utan kravet att korrelationerna ska följa varandra. Varje lag får sin egen korrelationsparameter korrelationsparameter.

Strukturer för kovariansmatrisen $\boldsymbol{R}$ - Toeplitz

$$\boldsymbol{R} = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots \rho_{n_j-1} \\ \rho_1 & 1 & \rho_1 & \dots \rho_{n_j-2} \\ \vdots & \vdots & \vdots & \\ \rho_{n_j-1} & \rho_{n_j-2} & \rho_{n_j-3} & \dots 1 \end{pmatrix}$$

- Samma varians men olika korrelationer mellan tidpunkter.

Det finns fler kovariansmatriser men dessa blir lite överkurs.

R: Övriga kovariansstrukturer

Strukturer för kovariansmatrisen $\boldsymbol{R}$ - Övriga

Spatial Power	SP(POW)(c)	$\sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \rho^{d_{14}} \\ \rho^{d_{21}} & 1 & \rho^{d_{23}} & \rho^{d_{24}} \\ \rho^{d_{31}} & \rho^{d_{32}} & 1 & \rho^{d_{34}} \\ \rho^{d_{41}} & \rho^{d_{42}} & \rho^{d_{43}} & 1 \end{bmatrix}$
Heterogeneous AR(1)	ARH(1)	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix}$
First-Order Autoregressive Moving-Average	ARMA(1,1)	$\sigma^2 \begin{bmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{bmatrix}$
Heterogeneous CS	CSH	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix}$
First-Order Factor Analytic	FA(1)	$\begin{bmatrix} \lambda_1^2+d_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 \\ \lambda_2\lambda_1 & \lambda_2^2+d_2 & \lambda_2\lambda_3 & \lambda_2\lambda_4 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2+d_3 & \lambda_3\lambda_4 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4^2+d_4 \end{bmatrix}$
Huynh-Feldt	HF	$\begin{bmatrix} \sigma_1^2 & \frac{\sigma_1^2+\sigma_2^2}{2}-\lambda & \frac{\sigma_1^2+\sigma_3^2}{2}-\lambda \\ \frac{\sigma_2^2+\sigma_1^2}{2}-\lambda & \sigma_2^2 & \frac{\sigma_2^2+\sigma_3^2}{2}-\lambda \\ \frac{\sigma_3^2+\sigma_1^2}{2}-\lambda & \frac{\sigma_3^2+\sigma_2^2}{2}-\lambda & \sigma_3^2 \end{bmatrix}$
Heterogeneous Toeplitz	TOEPH	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_2 & \sigma_1\sigma_4\rho_3 \\ \sigma_2\sigma_1\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_1 & \sigma_2\sigma_4\rho_2 \\ \sigma_3\sigma_1\rho_2 & \sigma_3\sigma_2\rho_1 & \sigma_3^2 & \sigma_3\sigma_4\rho_1 \\ \sigma_4\sigma_1\rho_3 & \sigma_4\sigma_2\rho_2 & \sigma_4\sigma_3\rho_1 & \sigma_4^2 \end{bmatrix}$
Unstructured Correlations	UNR	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{21} & \sigma_1\sigma_3\rho_{31} & \sigma_1\sigma_4\rho_{41} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \sigma_2\sigma_3\rho_{32} & \sigma_2\sigma_4\rho_{42} \\ \sigma_3\sigma_1\rho_{31} & \sigma_3\sigma_2\rho_{32} & \sigma_3^2 & \sigma_3\sigma_4\rho_{43} \\ \sigma_4\sigma_1\rho_{41} & \sigma_4\sigma_2\rho_{42} & \sigma_4\sigma_3\rho_{43} & \sigma_4^2 \end{bmatrix}$

Hur man väljer rätt struktur på kovariansmatrisen R

För att välja denna kan man använda ett utvärderingsmått som är antingen AIC eller BIC. Modeller med stora värden på log-likelihoodfunktionens maximum dvs med små värden på $-\log(L)$ samt få skattade parametrar är att föredra.

Ett jämförelsemått som tar hänsyn till antal skattade parametrar i kovariansmatrisen (q) är AIC:

$$\text{AIC} = -2\log(L) + 2q$$

Sedan väljer man modellen med lägst AIC/BIC.

Icke linjär data: modell med tidspolynom

I det fall man har icke linjär data kan man använda tidspolynom. Detta är svårt att se om man inte visualiserar datamaterialet därför är det viktigt att plotta all data för

att kunna konstantera om det är värt att modellera med polynom. I så fall görs det genom:

Modell med tidspolynom

Enkel två-nivåmodell med tidspolynom

- Hierarkisk form:
  - Nivå 1: $Y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_{2j} t_{ij}^2 + \cdots + \varepsilon_{ij}$
  - Nivå 2: $\beta_{0j} = \beta_0 + u_{0j}$
    $\beta_{1j} = \beta_1 + u_{1j}$
    $\beta_{2j} = \beta_2 + u_{2j}$
    $\vdots$
- Kombinerad form:
  - $Y_{ij} = (\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \dots) + (u_{0j} + u_{1j} t_{ij} + u_{2j} t_{ij}^2 + \cdots + \varepsilon_{ij})$
- $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{R})$
- $$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ \vdots \\ u_{n_j j} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} & \sigma_{u02} & \dots & \sigma_{u0n_j} \\ \sigma_{u10} & \sigma_{u1}^2 & \dots & & \vdots \\ \sigma_{u20} & \sigma_{u21} & \dots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \sigma_{un_j0} & \sigma_{un_j1} & \sigma_{un_j2} & \dots & \sigma_{un_jn_j}^2 \end{bmatrix} \right)$$

Modell med oberoende variabler

Två-nivåmodell med två oberoende variabler x_{j1} x_{j2} och tidspolynom $t_{ij} + t_{ij}^2$

- Hierarkisk form:
  - Nivå 1: $Y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_{2j} t_{ij}^2 + \varepsilon_{ij}$
  - Nivå 2: $\beta_{0j} = \beta_0 + \beta_3 x_{j1} + \beta_4 x_{j2} + u_{0j}$
    $\beta_{1j} = \beta_1 + \beta_5 x_{j2} + u_{1j}$
    $\beta_{2j} = \beta_2 + \beta_6 x_{j2} + u_{2j}$

Modell med oberoende variabler

- Kombinerad form:
  - $Y_{ij} = (\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 x_{j1} + \beta_4 x_{j2} + \beta_5 x_{j2} t_{ij} + \beta_6 x_{j2} t_{ij}^2) + (u_{0j} + u_{1j} t_{ij} + u_{2j} t_{ij}^2 + \varepsilon_{ij})$

- $\boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \boldsymbol{R})$

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

Flernivå mixed modell

Vi har följande datamaterial vilket består av flera nivåer

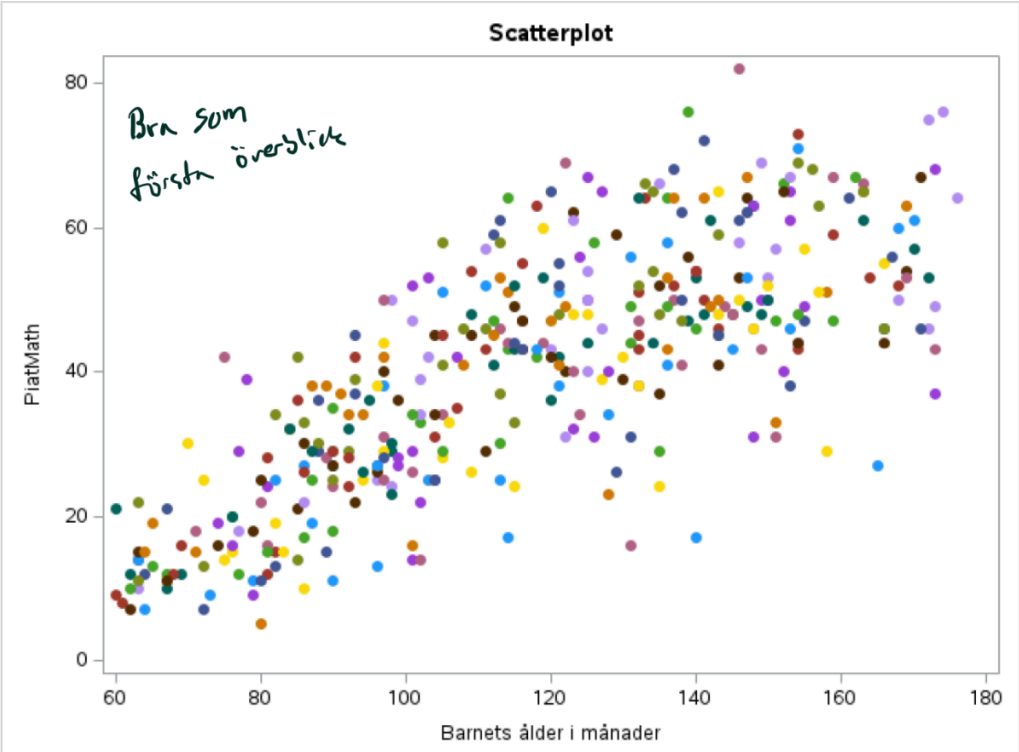
Exempel: NLSY79

Kolumnerna ger: observation, individ, familj, etnicitet, kön, födelseår, födelseordning, mors ålder vid barnets födsel, Piat Math, Piat Math åldersstandardiserat, ålder (månader). Piat Math mäter prestationer i matematik. Den består av 84 flervalsfrågor med ökande svårighetsgrad.

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

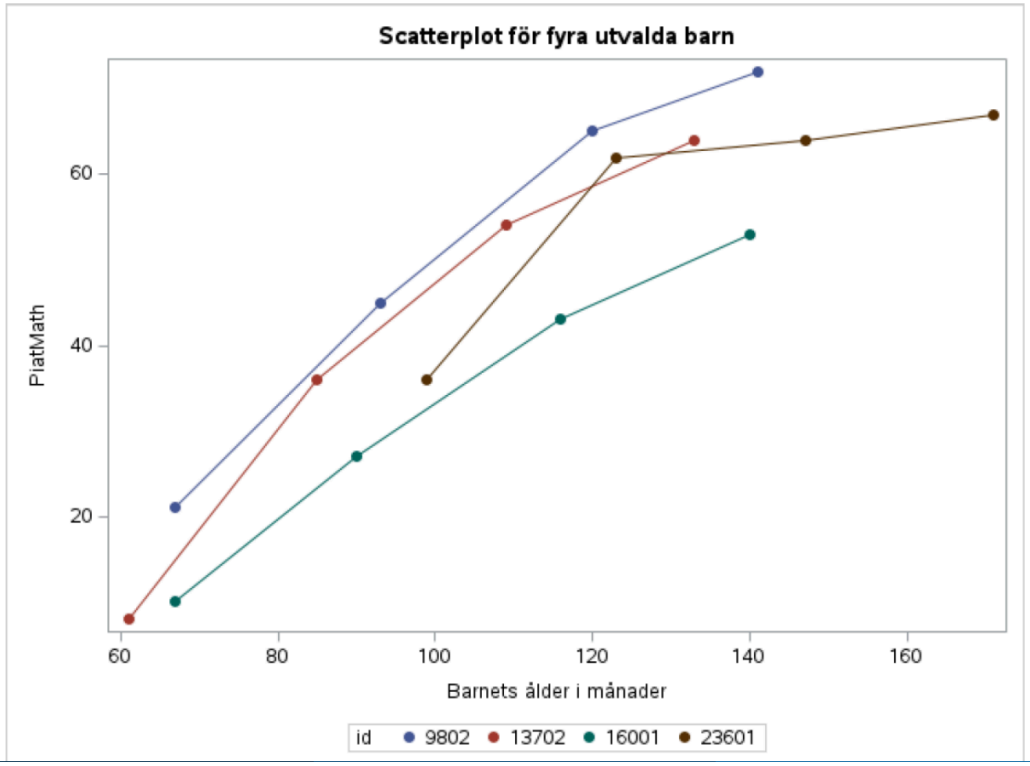
Vi visualiserar datamaterialet och noterar att utvecklingen inte verkar vara linjär.

Exempel: NLSY79 Scatter plot



Vi kollar lite närmare på några fåtal observationer och bedömmer utifrån dessa om det är värt att modellera med polynom.

Exempel: NLSY79 Scatter plot fyra individer



Modellformuleringen behöver ändras någorlunda vilket nu blir:

Flernivå mixed model

Nivå 1: $m_{ij} = \beta_{0j} + \beta_{1j} a_{ij} + \varepsilon_{ij}$
Nivå 2: $\beta_{0j} = \beta_0 + u_{0j}$
$\beta_{1j} = \beta_1 + u_{1j}$

Kombinerad form:
$m_{ij} = (\beta_0 + \beta_1 a_{ij}) + (u_{0j} + u_{1j} a_{ij} + \varepsilon_{ij})$
$\varepsilon_i \sim N(0, \boldsymbol{R})$
$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix} \right)$$