Hemtentamen 2 - Longitudinell data

732G34 Statistiska metoder för komplex data

Hampus Beijer

21 november 2025

1 Syfte

2 Datmaterial

Obs	id	idmom	race	sex	birthyear	birthorder	agemom	a	m	mz	time
1	623802	6238	2	1	1990	2	28	72	16	103	0
2	623802	6238	2	1	1990	2	28	96	25	93	1
3	623802	6238	2	1	1990	2	28	122	48	108	2
4	623802	6238	2	1	1990	2	28	146	57	109	3
5	411002	4110	3	1	1989	2	27	85	21	99	0
6	411002	4110	3	1	1989	2	27	107	41	105	1
7	411002	4110	3	1	1989	2	27	136	47	100	2
8	411002	4110	3	1	1989	2	27	157	53	100	3
9	849203	8492	2	1	1988	3	29	93	22	92	0
10	849203	8492	2	1	1988	3	29	116	35	89	1
11	849203	8492	2	1	1988	3	29	144	43	89	2
12	849203	8492	2	1	1988	3	29	166	53	97	3
13	380601	3806	3	1	1988	1	30	90	31	106	0
14	380601	3806	3	1	1988	1	30	114	51	117	1
15	380601	3806	3	1	1988	1	30	141	52	104	2
16	380601	3806	3	1	1988	1	30	165	59	102	3
17	343003	3430	3	2	1989	3	29	87	42	120	0
18	343003	3430	3	2	1989	3	29	110	47	113	1
19	343003	3430	3	2	1989	3	29	135	51	106	2
20	343003	3430	3	2	1989	3	29	159	67	116	3

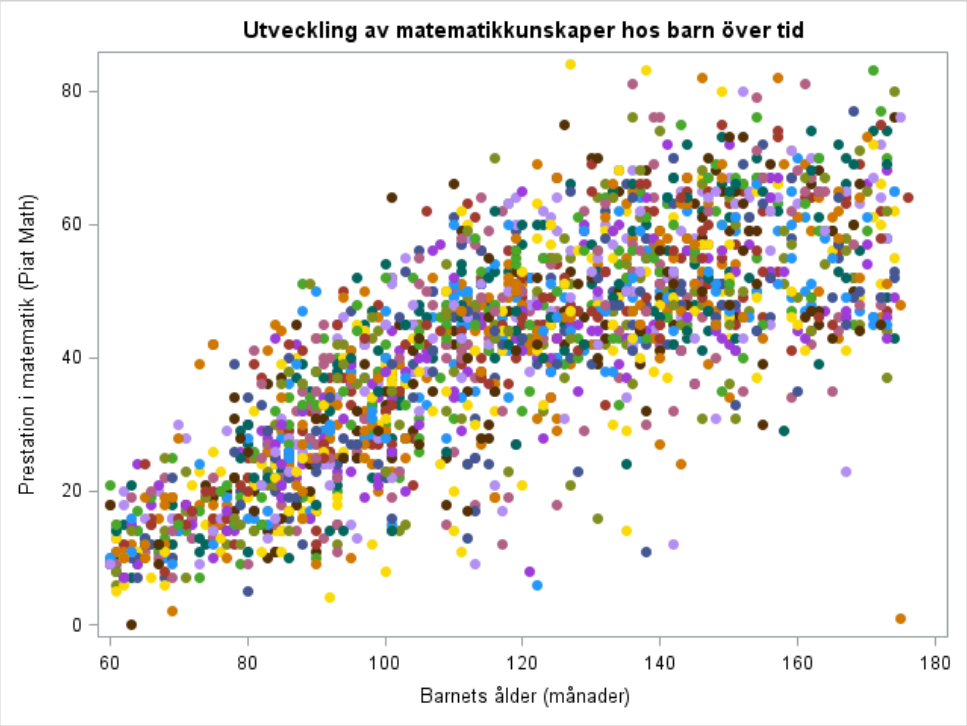
Tabell 1: De första 20 observationerna av datamaterialet.

3 Uppgift 1

3.1 Deluppgift 1a

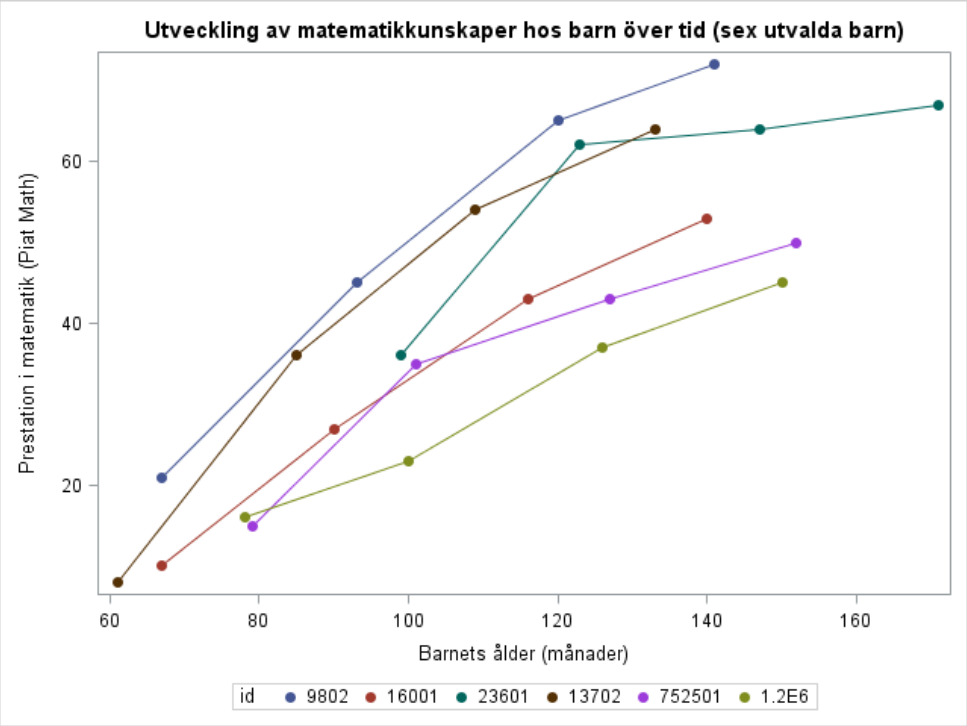
Gör lämpliga plottar av datamaterialet och beskriv eventuella mönster. (1p)

I figur 1 presenteras en graf över barns utveckling i matematik. I grafen presenteras barns ålder i månader på x-axeln och prestation i matematik (Piat Math) på y-axeln. Utifrån datamaterialet ser man att det verkar finnas en positiv korrelation, alltså, ju äldre ett barn blir desto bättre presterar den i matematik. Man kan notera att utvecklingsmönstret inte verkar linjärt, där man ser efter 120 månader börjar barns utveckling saktas ned. På grund av det är många barn i detta datamaterial presenteras istället ett urval av barn för att kolla närmare på datamaterialet.



Figur 1: Diagrammet visar utvecklingen av barns prestation i matematik. Barns ålder i månader visas på x-axeln och prestationer i matematik på y-axeln.

I figur 2 visas ett urval av sex barns utveckling i matematik. I denan graf ser man tydligare att utvecklingen verkar icke-linjär. Det kan därför vara av intresse att modellera med polynom i senare delar av uppgifterna.



Figur 2: Diagrammet visar utvecklingen av sex barns prestation i matematik. Barns ålder i månader visas på x-axeln och prestationer i matematik på y-axeln. Linjerna tyder på icke-linjäritet.

3.2 Deluppgift 1b

Skatta en lämplig två-nivå modell med matematikresultat som beroende variabel och ålder i månader som oberoende tidsvariabel samt eventuella lämpliga tidspolynom. Antag ingen kovariansstruktur i R-matrisen förutom konstant varians. Motivera dina val noggrant och tolka resultaten. (3p)

I denna deluppgift skapas en flernivå mixed model. Jag börjar med att visa ekvationsformuleringarna av nivå 1 och 2, kombineringen av nivåerna, modellens antaganden, därefter tolkas modellen.

$$\textbf{Nivå 1 (inom individer) :}\ m_{ij} = \beta_{0j} + \beta_{1j}a_{ij} + \epsilon_{ij} \tag{1}$$

$$\textbf{Nivå 2 (mellan individer) :}\ \beta_{0j} = \beta_0 + u_{0j} \quad \beta_{1j} = \beta_1 + u_{1j} \tag{2}$$

$$\textbf{Kombinerad form :}\ m_{ij} = \underbrace{(\beta_0 + \beta_1 a_{ij})}_{\text{Fixed}} + \underbrace{(u_{0j} + u_{1j}a_{ij})}_{\text{slumpmässig}} + \epsilon_{ij} \tag{3}$$

$$\textbf{Modellens antaganden :}\ \epsilon \sim N(0, \boldsymbol{R}) \quad \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix} \right) \tag{4}$$

Ovan kan ekvationsformuleringarna ses, där ekvationsformuleringarna använder datamaterialets variabelnamn. Det vill säga, m står för prestationen i matematik och a står för barnets ålder i månader. Modellformuleringen för nivå 1 modellerar hur barnets reslutat vid varje mättillfälle beror på åldern, där i är tidpunkten och j är barnet. β_{0j} är det individuella interceptet för varje barn, β_{1j} är varje barns lutning, alltså hur mycket reslutatet förändras varje månad. ϵ_{ij} är residualen.

För nivå 2 modelleras varje barns utveckling, alltså lutningen och interceptet för varje barn. β_{0j} är barnets (j) intercept och $\beta 1j$ lutning. Det bör förtydligas här att varje barn har sin egen lutning (vilket kan ses i figur 1 och 2 när datamaterialet visades). För interceptet och lutningen är β_1 och β_0 fixa effekter medan u_{0j} och u_{1j} är slumpmässiga effekter (hur varje barn j avviker från genomsnittet).

Den tredje ekvationsformuleringen, den kombinerade formen är en ihopskrivning av nivå 1 och 2. Här kan det ses det tydligare vilka parametrar som är fixa och slumpmässiga. Den första delen $\beta_0 + \beta_1 a_{ij}$ visar det genomsnittliga interceptet och lutningen, det vill säga, vart startnivån för prestation i matematik vid en viss ålder bör ligga och den genomsnittliga ökningen för ett barn varje månad. Den slumpmässiga delen $u_{0j} + u_{1j}a_{ij}$ visar hur varje barns utveckling skiljer sig åt. Denna del tillåter att varje barn har ett individuellt intercept och lutning, det vill säga man möjliggör att kunna analysera hur varje barn skiljer sig åt från genomsnittet. Den sista delen är ϵ_{ij} vilket är residualerna (tillfälliga variationer).

I modellen antas att residualerna följer en normalfördelning med väntevärde 0 och kovarians $\mathbf{R}$. I denna deluppgift antas ingen kovariansstruktur vilket medför att den ser ut på följande sätt: sim $N(0, \sigma^2)$, alltså väntevärde 0 och konstant varians. De slumpmässiga effekterna u_{0j} och u_{1j} antas följa en normalfördelning med väntevärdesvektor 0 och en kovariansmatris (oftasts betecknad $\mathbf{G}$). För denna kovariansmatris ser man varianserna på diagonalen för σ^2_{u0} (slumpmässiga interceptet) och σ^2_{u1} (slumpmässiga lutningen). Utanför diagonalerna ses kovarianserna mellan de slumpmässiga effekterna. Med det sagt berättar kovariansmatrisen hur effekterna varierar och sammanhänger.

3.2.1 Modellanpassning

I modellbeskrivningen ovan presenterades att varje barn fick en slumpmässig lutning, men detta utgjorde ett problem. När man inkluderade en slumpmässig lutning kunde blev matrisen $\mathbf{G}$ icke positivt definite. Detta problem löstes genom att ta bort den slumpmässiga lutningen. Att ta bort den slumpmässiga

lutningen innebär att man nu antar att alla barns utveckling är densamma att det ökar för att sedan avta, vilket man faktiskt ser i figur 1 och 2. Modellformuleringen ändras nu till:

$$\textbf{Nivå 1 (inom individer):}\ m_{ij} = \beta_{0j} + \beta_1 a_{ij} + \epsilon_{ij} \tag{5}$$

$$\textbf{Nivå 2 (mellan individer):}\ \beta_{0j} = \beta_0 + u_{0j} \tag{6}$$

$$\textbf{Kombinerad form:}\ m_{ij} = \underbrace{(\beta_0 + \beta_1 a_{ij})}_{\text{fixed}} + \underbrace{u_{0j}}_{\text{slumpmässig intercept}} + \epsilon_{ij} \tag{7}$$

$$\textbf{Modellens antaganden:}\quad \epsilon \sim N(0, \boldsymbol{R}) \quad \text{och} \quad u_{0j} \sim N\left(0, \sigma^2_{u0}\right) \tag{8}$$

Tolkningen av denna modell är, i princip, samma som tidigare. Skillnaden ligger i att den slumpmässiga lutningen är borta. På grund av tidsbrist tolkas inte modellen i detalj. Men kort sagt tillåter denna modell att intercepten varierar mellan barn (u_{0j}), medan barn även får ett genomsnittligt intercept med en gemensam lutning som nu antas vara lika för alla barn.

Vidare modellerades tre två-nivå modeller en modell med grad 1 polynom (enligt ekvationsformuleringen ovan), andragradspolynom modell, och en tredjegradspolynom modell. Anledningen till varför modeller med polynom skapades var på grund av datans icke-linjäritet. I tabell 2 presenteras modellernas olika AIC, vilket visar att en modell av grad 2 kan vara lämpligre än grad 1 och 3.

Modell	Polynom (grad)	AIC
Modell 1	1	14329.8
Modell 2	2	14043.5
Modell 3	3	14054.9

Tabell 2: Mixed effekt modeller med AIC.

3.2.2 Modell med grad två polynom

Den bäst presterande modellen (utifrån AIC) är den modell med ett polynom. På grund av att ett polynom läggs till i modellen förändras nu modellekvationen till:

$$\textbf{Nivå 1 (inom barn):}\quad m_{ij} = \beta_{0j} + \beta_1 a_{ij} + \beta_2 a^2_{ij} + \epsilon_{ij} \tag{9}$$

$$\textbf{Nivå 2 (mellan barn):}\quad \beta_{0j} = \beta_0 + u_{0j} \tag{10}$$

$$\textbf{Kombinerad form:}\quad m_{ij} = \underbrace{\left(\beta_0 + \beta_1 a_{ij} + \beta_2 a^2_{ij}\right)}_{\text{fixa effekter}} + \underbrace{u_{0j}}_{\text{slumpmässigt intercept}} + \epsilon_{ij} \tag{11}$$

Tabell 3 presenterar parameterskattningarna för kovariansmatrisen. På grund av att endast ett slumpmässigt intercept finns bara en skattning. $UN(1,1) = 60.29$, vilket är variansen för det slumpmässiga interceptet. Variansen är mycket hög vilket tyder på att barns startnivå i matematik skiljer stort mellan barn. Residualen är 39.73 vilket presenterar inom-individ variationen. Residualen säger hur mycket variation över tiden som modellen inte lyckas fånga, vilket är relativt högt.

Cov Parm	Subject	Estimate	Std. Error	Z Värde	Pr Z
UN(1,1)	id	60.2948	4.4619	13.51	0.0001
Residual		39.7318	1.4518	27.37	0.0001

Tabell 3: Parameterskattningar för kovariansmatrisen (modell med enbart slumpmässigt intercept).

I tabell 4 presenteras utvärderingsmått för modellen. Här kan man notera AIC, där man vill välja den modell som har lägst AIC. Kort så straffar AIC modeller baserat på modellens komplexitet. Hur som heslt bör man ta detta med en nypa salt. Endast för att andragradspolynom modellen har bäst AIC innebär detta inte att detta alltid kommer vara den bästa.

Statistiska	Värde
-2 Res Log Likelihood	14039.5
AIC	14043.5
AICC	14043.5
BIC	14051.9

Tabell 4: Utvärderingsmått för modellen.

I tabell 5 presenteras ett likelihood ratio test (χ^2) vilket visar att modellen med ett andragradspolynom är bättre än en modell med endast ett intercept.

DF	Chi-Square	Pr > ChiSq
1	865.90	0.0001

Tabell 5: Null Model Likelihood Ratio Test.

I tabell 6 presenteras de fasta effekterna. Det bör noteras att variabel a har centrerats för att underlätta tolkning. Interceptet är 44.13 vilket innebär att genomsnittliga nivån för ett barns prestation i matematik är 44 poäng. Variabeln Ca vilket är centrerade a visar en skattning på 0.4504, alltså i genomsnitt ökar barns prestation i matematik med 0.45 poäng vid varje ökning i tiden. Andragradspolynomet Ca^2 visar en skattning på -0.00347 vilket visar att utvecklingen avtar över tid. Detta verkar mycket rimiligt eftersom man kan se detta mönster i figur 1 och 2. Alla variabler är signifikanta.

Effect	Estimate	Std. Error	DF	t Värde	Pr > |t|
Intercept	44.1283	0.4009	499	110.07	0.0001
Ca	0.4504	0.005089	1498	88.50	0.0001
Ca*Ca	-0.00347	0.000159	1498	-21.86	0.0001

Tabell 6: Skattningar av fasta effekter för den kvadratiska modellen.

Tabell 7 presenterar typ 3 test för fixa effekter vilket innebär att man testar för att se om varje fast effekt bidrar signifikant till modellen. För både Ca och $Ca * Ca$ noteras det att dessa är signifkanta vilket innebär att man utvecklingen över tid är linjärt signifkant och kvadratiskt signifkant. Detta innebär i praktiken att barn utvecklas i matematik över tid, där man ser att utvecklingen inte endast är linjär utan man ser ett icke-linjärt mönster.

Effect	Num DF	Den DF	F Värde	Pr > F
Ca	1	1498	7831.42	0.0001
Ca*Ca	1	1498	477.65	0.0001

Tabell 7: Type 3 Tests of Fixed Effects för den kvadratiska modellen.

3.3 Deluppgift 1c

Utgår från modellen i 1 b) och testa om det finns någon/några bakgrundsvariabler som har effekt på matematikresultaten. Motivera dina val noggrant och tolka resultaten. (3p)

I denna deluppgift läggs de förklarande variablerna till i modellen. Först läggs *race* till i modellen (tabell 8), där man kan notera att AIC är 13947.6. AIC minskar alltså när race läggs till i modellen. Detta kan troligtvis förklaras av socioekonomiska faktorer.

Modell	**Inkluderade variabler**	**Signifikanta variabler**	**AIC**
Basmodell (utan race)	Intercept, Ca, Ca^2	Ca (p=0.0001), Ca^2 (p=0.0001)	14043.5
Modell med race	Intercept, Ca, Ca^2, race	Ca (p=0.0001), Ca^2 (p=0.0001), race (p=0.0001)	13947.6

Tabell 8: Jämförelse mellan basmodellen och modellen med bakgrundsvariabeln *race*.

Vidare läggs *sex*, *CAgemom*, och *CBirthyear* till i modellen där man kan notera i tabell 9 att dessa variabler blir icke-signifikanta. Därför exkluderas dessa, men modellen med race läggs till på grund av att AIC blir mindre och variabeln är signifkant.

Modell	**Inkluderade variabler**	**Signifikans (p-värden)**	**AIC**
Modell 1: Basmodell + race	Ca, Ca^2, race	Ca (p=0.0001), Ca^2 (p=0.0001), race (p=0.0001)	13947.6
Modell 2: Basmodell + race + sex	Ca, Ca^2, race, sex	Ca (p=0.0001), Ca^2 (p=0.0001), race (p=0.0001), sex (p=0.5355)	13946.1
Modell 3: Basmodell + race + CAge_mom	Ca, Ca^2, race, CAge_mom	Ca (p=0.0001), Ca^2 (p=0.0001), race (p=0.0001), CAge_mom (p=0.0700)	13946.3
Modell 4: Basmodell + race + Cbirthyear	Ca, Ca^2, race, Cbirthyear	Ca (p=0.0001), Ca^2 (p=0.0001), race (p=0.0001), Cbirthyear (p=0.1282)	13945.6

Tabell 9: Jämförelse av modeller med olika bakgrundsvariabler.

I tabell 10 presenteras kovariansmatrisen där man kan notera att $\text{UN}(1,1) = 48.63$ vilket är mindre för modellen utan variabeln *race*. Detta tolkas som barns startnivå i prestation i matematik kan förklaras av race. Med andra ord kan bidrar racetill att minska den oförklarade mellan individ variationen, vilket innebär att modellen kan bättre förklara skillnader mellan barns prestationer i matematik.

Cov Parm	Subject	Estimate	Std. Error	Z Värde	Pr Z
UN(1,1)	id	48.6286	3.7331	13.03	0.0001
Residual		39.7302	1.4517	27.37	0.0001

Tabell 10: Parameterskattningar för kovariansmatrisen (modell med race).

I tabell 11 visas utvärderingsmått, där man bland annat kan notera att modellens AIC minskar med cirka 100 enheter jämfört med modellen utan race.

Statistiska	Värde
-2 Res Log Likelihood	13943.6
AIC	13947.6
AICC	13947.6
BIC	13956.0

Tabell 11: Utvärderingsmått för modellen med race.

I tabell 12 presenteras ett likelihood ratio test vilket säger att en modell med race är signifkant bättre än en modell utan.

DF	Chi-Square	Pr > ChiSq
1	708.39	0.0001

Tabell 12: Null Model Likelihood Ratio Test för modellen med race.

I tabell 13 presenteras parameterskattningarna för de fixa effekterna. Interceptet är 46.93, alltså den genomsnittliga nivån av ett barns prestation i matematik är 49.6 poäng. Ca och Ca^2 är 0.4508 och -0.00347 vilket visar att barns utveckling i matematik ökar men efter ett tag börjar ökningen avta. För *race* är $race = nonhispanic/nonblack(3)$ referenskategorin, där skattningarna för $race = hispanic(1)$ och $race = black(2)$ är -5.36 respektive -7.66. Det vill säga att "hispanicbarn presterar i genomsnitt 5.36 poäng sämre än non-hispanicöch non-blacks", och barn som är blackpresterar i genomsnitt 7.66 poäng sämre än non-hispanicsöch non-blacks". Alla skattade effekter är mycket signifkanta. Med detta sagt kan man notera att ras verkar spela en stor roll i prestation i matematik för barn.

I tabell 14 presenterar om de fixa effekterna faktiskt bidrar till modellen. Det kan noteras att alla p-värden är mycket signifkanta. Det vill säga, tiden påverkar ett barns prestation i matematik (Ca), men samtidigt

Effect	Estimate	Std. Error	DF	t Värde	Pr > $\|t\|$
Intercept	46.9292	0.4670	497	100.50	0.0001
Ca	0.4508	0.005076	1498	88.81	0.0001
Ca*Ca	-0.00347	0.000159	1498	-21.85	0.0001
race 1	-5.3600	1.0171	497	-5.27	0.0001
race 2	-7.6590	0.7969	497	-9.61	0.0001
race 3	0	.	.	.	.

Tabell 13: Skattningar av fasta effekter för modellen med race.

kan man noter att denna utveckling inte är linjär (Ca^2), och man ser en skilland i barns prestation i mateamtik över tiden baserat på deras ras (race).

Effect	Num DF	Den DF	F Värde	Pr > F
Ca	1	1498	7887.53	0.0001
Ca*Ca	1	1498	477.36	0.0001
race	2	497	50.73	0.0001

Tabell 14: Type 3 Tests of Fixed Effects för modellen med race.

3.4 Deluppgift 1d

Utgå från modellen i 1 c) och undersök om du behöver någon kovariansstruktur i R-matrisen. Resonera över hur korrelation mellan tidpunkter skulle kunna se ut för den här typen av datamaterial och välj några strukturer för att jämföra. Kan du se om någon struktur verkar passa bättre? Motivera dina val noggrant och tolka resultaten. (2p)

För detta datamaterial har det man mätt barns prestation i matematik över fyra tidspunkter. Det är mycket rimligt att anta positiv autokorrelation, alltså att ett barns prestation i matematik utvecklas över tiden. Det är även rimligt att anta att korrelationen är starkare när observationerna är nära varnadra t.ex tidpunkt $1-2$ jämfört med tidpunkt $1-4$. med detta i åtanke kan en AR(1), alternativt SP(POW) struktur på kovariansmatrisen $\mathbf{R}$ vara lämplig. Hur som helst kommer jag att testa alla strukturer för att kunna jämföra om en struktur är bättre än en annan.

3.4.1 CS (Compound Symmetry)

Compound symmetry är en av de enklaste strukturerna för kovariansmatrisen $\mathbf{R}$, som är lämplig i det fall när man antar att feltermerna är korrelerade mellan tidpunkter inom individer. När compound symmetry strukturen användes för modellen i deluppgift 1c) konvergerade modellen, men matrisen med partiella andraderivator (Hessianen) är inte positive definite vilket medför att modellen är numeriskt instabil. Konsekvent innebär detta att denna struktur på matrisen $\mathbf{R}$ inte är lämplig. Alltså, denna struktur är inte lämplig för detta datamaterial.

3.4.2 UN (Unstructured)

En UN (Unstructured) struktur testades för modellen i deluppgift c). Denna struktur är lämplig när datamaterialet har få mättillfällen på grund av att det skattas en unik varians och kovarians för varje tidskombination, konsekvenst, fler blir det parametrar att skatta. För det givna datamaterialet lyckas modellen konvergera, men åter igen är hessian inte positivt definite. Detta innebär att denna struktur ej är lämplig för det givna datamaterialet.

3.4.3 AR(1)

En autoregressive struktur på kovariansmatrisen är mycket bra för detta datamaterialet eftersom man kan anta positiv autkorrelation (vilket man ser i figur 1), det vill säga, om ett barn har ett högt värde vid en tidpunkt är det mycket troligt att tidpunkten därefter kommer vara mycket lika. I sin tur innebär detta även att korrelationen mellan t.ex tidpunkt 1 och 2 är högre än tidpunkt 1 och 4, vilket är en bra motivering till varför AR(1) faktiskt bör användas för detta datamaterial. Ett antagande som måste uppfyllas för AR(1) struktur på kovariansmatrisen är att tidsintervallen ska vara jämna; detta uppfylls.

En modell med denna struktur skattades. Modellen lyckades konvergera och det blev inget problem med att beräkna hessianen. Reslutaten presenteras i tabell 15-18. I tabell 15 presenteras parameterskattningarna för kovariansmatrisen vilket visar att den skattade variansen för det slumpmässiga interceptet är 42.90 och signifikant. Det slumpmässiga interceptet tyder på att barn skiljer sig (tydligt) mellan barn. Den skattade AR(1) parametern är 0.196 och signifkant, vilket visar korrelationen mellan residualerna.Att skattningen är $\rho = 0.196$ innebär att korrelationen är svag, men positivt korrelad. Alltså, mätningar som är nära varandra är något lika. Residualvariansen är 44.90 som beskriver den oförklarade variationen (som modellen inte lyckas fånga).

I tabell 16 presenteras utvärderingsmått för modellen. Man kan notera att AIC har minskat med 21.2 enheter jämfört med modellen utan en kovariansstruktur. En modell med AR(1) som kovariansstruktur är bättre än en modell utan kovariansstruktur.

I tabell 17 presenteras tester för de skattade (fixa) effekterna, som visar att alla effekter är statistiskt signifkanta. Det vill säga, det finns en linjär effekt (Ca), en icke linjär effekt (Ca*Ca), och en effekt av vilken ras barnet tillhör.

Tabell 18 presenteras den skattade korrelationsmatrisen för AR(1), vilket visar att korrelationen minskar när tidsavståndet ökar. Korrelationen vid lag 1 är 0.196, 0.038 vid lag 2, 0.0038 vid lag 3, och minst är korrelationen vid lag 4 med 0.00075. Det vill säga, AR(1) strukturen verkar mycket lämplig.

Cov Parm	Subject	Estimate	Std. Error	Z Värde	Pr Z
Intercept	id	42.9027	4.0421	10.61	0.0001
AR(1)	id	0.1959	0.04265	4.59	0.0001
Residual		44.8959	2.3708	18.94	0.0001

Tabell 15: Parameterskattningar för kovariansmatrisen (AR(1)-struktur).

Statistiska	Värde
-2 Res Log Likelihood	13920.4
AIC	13926.4
AICC	13926.4
BIC	13939.0

Tabell 16: Utvärderingsmått för modellen med AR(1)-struktur.

Effect	Num DF	Den DF	F Värde	Pr > F
Ca	1	1498	6568.16	0.0001
Ca*Ca	1	1498	431.25	0.0001
race	2	497	51.52	0.0001

Tabell 17: Type 3 Tests of Fixed Effects för AR(1)-modellen.

Row	**Col1**	**Col2**	**Col3**	**Col4**
1	1.0000	0.1959	0.03836	0.007514
2	0.1959	1.0000	0.1959	0.03836
3	0.03836	0.1959	1.0000	0.1959
4	0.007514	0.03836	0.1959	1.0000

Tabell 18: Skattad R-korrelationsmatris för AR(1)-strukturen

3.4.4 SP(POW)

SP(POW) är en kovariansstruktur som används när tidspunkterna i ens datamaterial har olika tidsavstånd, men om tidpunkterna är jämt fördelade är SP(POW) ekvivalent med AR(1). När en modell med denna struktur skattades visades det att AIC är densamma som för AR(1), vilket mest troligtvis beror på datamaterialet. På grund av att tidsavstånden i detta datamaterial följer en jämn fördelning, så är SP(POW) eqvivalent med en AR(1). Skattningarna presenteras i tabell 19-21, men tolkas inte på grund av dessa är nästan samma för AR(1)-strukturen.

Cov Parm	Subject	Estimate	Std. Error	Z Värde	Pr Z
Intercept	id	42.9027	4.0421	10.61	0.0001
SP(POW)	id	0.1959	0.04265	4.59	0.0001
Residual		44.8959	2.3708	18.94	0.0001

Tabell 19: Parameterskattningar för kovariansstrukturen med Spatial Power (SP(POW)).

3.4.5 Toeplitz

Toeplitz struktur på kovariansmatrisen kan användas när man antar att ens data har samma varians vid alla tidpunkter, men att korrelationerna är olika mellan tidpunkterna. När denna struktur användes för

	Col1	Col2	Col3	Col4
Row 1	1.0000	0.1959	0.03836	0.007514
Row 2	0.1959	1.0000	0.1959	0.03836
Row 3	0.03836	0.1959	1.0000	0.1959
Row 4	0.007514	0.03836	0.1959	1.0000

Tabell 20: Skattas R korrelationsmatris (SP(POW)).

Statistiska	Värde
-2 Res Log Likelihood	13920.4
AIC	13926.4
AICC	13926.4
BIC	13939.0

Tabell 21: Utvärderingsmått för modellen med Spatial Power-struktur.

modellen i deluppgift 1c lyckades modellen konvergerade, men hessianen är inte positivt definivt. Det vill säga, denna kovariansstruktur anses inte lämplig för det givna datamaterialet.

3.4.6 Jämförelse av kovariansstrukturer

Det märktes när modellerna med de olika kovariansstrukturer att följande kovariansstrukturer inte är lämpliga för det valda datamaterialet:

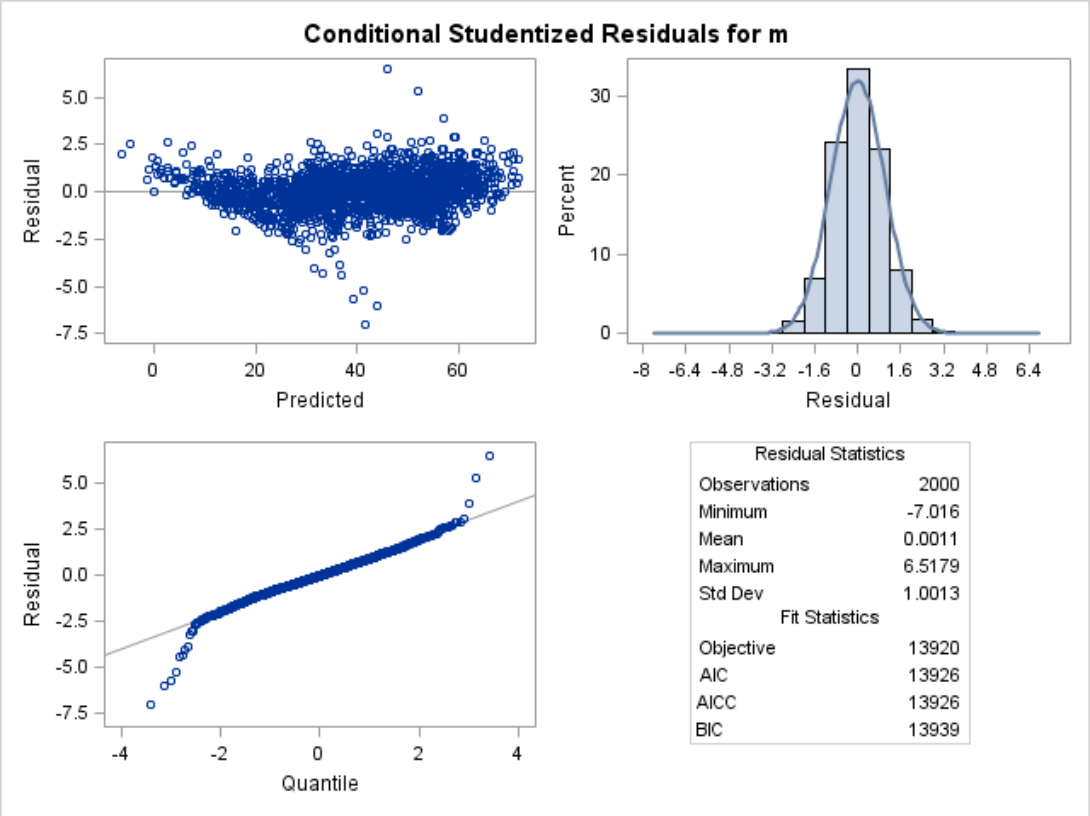
- CS (Compound Symmetry)
- UN (Unstructured)
- Toeplitz

problemet med dessa kovariansstrukturer är att modellens hessian inte är positivt definite, vilket tyder på att dessa inte är lämpliga för datamaterialet. De kovariansstrukturer som lyckas ge modeller utan problem med skattning är AR(1) och SP(POW). På grund av att tidsavstånden är jämna blev SP(POW) = AR(1). Den bästa strukturen är alltså att modellera med en AR(1) kovariansstruktur. Anledningen till varför denna struktur verkar vara den bästa är att AR(1) kräver specifika antaganden som denna datan uppfyllde.

3.5 Deluppgift 1e

Plotta residualerna för din slutgiltiga modell och avgör om antagandena är uppfyllda. Motivera ditt svar. (1p)

I figur 3 presenteras modellens (andragradspolynom) residualanalys. Längst upp i vänster hörn presenteras undersöker heteroskedasticitet. Man kan notera att de flesta av punkterna ligger spridda runt 0, med några få observationer som avviker vid predicted = 40. Men detta verkar inte vara särskilt allvarligt. Antagandet om heteroskedasticitet antas uppfyllas. Histogrammet (längst upp till höger) visar att residualerna verkar vara normalfördelade, det vill säga, antagandet om normalitet uppfylls. Hur som helst kan man notera att i QQ-plotten (längst ned i vänster hörn) att det finns några avvikelser i svansarna - med eftersom det endast är några få punkter kommer detta mest troligtvist inte att påverka modellen särskilt mycket. Med detta sagt verkar modellen lämplig utifrån residualanalysen.



Figur 3: Diagrammet visar utvecklingen av sex barns prestation i matematik. Barns ålder i månader visas på x-axeln och prestationer i matematik på y-axeln. Linjerna tyder på icke-linjäritet.

3.6 Bifogad SAS kod för uppgift 2

3.6.1 Deluppgift 1a

```
*Laddar in datamaterialet med vanliga datalines;
data data;
    infile datalines truncover missover;
    input
        id
        idmom
        race
        sex
        birthyear
        birthorder
        agemom
        a1 a2 a3 a4
        m1 m2 m3 m4
        mz1 mz2 mz3 mz4
        time1 time2 time3 time4
    ;
datalines;
*Observera detta är inte hela datmaterilaet kan inte bifoga 100 rader med data;
623802 6238 2 1 1990 2 28 72 96 122 146 16 25 48 57 103 93 108 109 0 1 2 3
;

*Transformerar datamaterialet (KOD bifogad från Jonas Bjermo);
        data wide_data; /* Blivande namn*/
```

```
set data; /* Nuvarande namn */
/* Skapa en ny kolumn "a" med värden från kolumnerna a1-a4
radvis */
array a_Värdes[4] a1-a4; /* Skapa en array för
variablerna a1 till a4 */
array m_Värdes[4] m1-m4; /* Skapa en array för
variablerna m1-m4 */
array mz_Värdes[4] mz1-mz4; /* Skapa en array för
variablerna mz1-mz4 */
array time_Värdes[4] time1-time4; /* Skapa en array för
variablerna time1-time4 */
do i = 1 to 4 ; /* Loop från 1 till 4*/
a = a_Värdes[i]; /* Tilldela värdet från
aktuell kolumn till "a" */
m = m_Värdes[i]; /* Tilldela värdet från
aktuell kolumn till "m" */
mz = mz_Värdes[i]; /* Tilldela värdet från
aktuell kolumn till "mz" */
time = time_Värdes[i]; /* Tilldela värdet från
aktuell kolumn till "time" */
output; /* Skriv ut varje rad med det nya värdet för
kolumnerna "a","m","mz", "time" */
end;
drop i a1 a2 a3 a4 m1 m2 m3 m4 mz1 mz2 mz3 mz4 time1 time2
time3 time4; /* Ta bort det som inte behövs*/
run;

*Kontrollerar strukturen på datamaterialet;
proc print data = wide_data(obs = 20);

*Centrera (grand mean) mina kontinuerliga variabler för att underlätta tolkning;

*Behöver skapa medelvärden för variabler;
proc means data = wide_data noprint;
        var agemom birthyear;
        output out = medel_data mean = medel_agemom medel_birthyear;
run;

*Nu centreras datamaterialet;
data centrerad_data;
if _n_ = 1 then set medel_data; *fixar medelvärden;
set wide_data;
CAge_mom = agemom -  medel_agemom;
Cbirthyear = birthyear - medel_birthyear;
        run;

*Kontrollerar strukturen på datamaterialet;
proc print data=centrerad_data;

*Plottar datamaterialet;
        proc sgplot data=centrerad_data;
scatter x=a y=m / group=id markerattrs=(symbol=circlefilled size=8);
title 'Utveckling av matematikkunskaper hos barn över tid';
xaxis label='Barnets ålder (månader)';
yaxis label='Prestation i matematik (Piat Math)';
run;

*plottar specifika observationer;
proc sgplot data=centrerad_data;
scatter x=a y=m / group=id markerattrs=(symbol=circlefilled);
series x=a y=m / group=id lineattrs=(pattern=solid);
where id in (9802, 13702, 16001,23601,752501,1201802);
title 'Utveckling av matematikkunskaper för sex utvalda barn';
```

```
xaxis label='Barnets ålder (månader)';
yaxis label='Prestation i matematik (Piat Math)';
run;
```

3.6.2 Deluppgift 1b

```
*Modellerar utan polynom;
proc mixed data =centrerad_data covtest;
class id;
model m=Ca / solution ddfm = bw;
random intercept Ca/ subject = id type = un gcorr;
run;

*Testar modellera med andragradspolynom;
proc mixed data=centrerad_data covtest;
    class id;
    model m = Ca Ca*Ca / solution ddfm=bw;
    random intercept / subject=id type=un gcorr;
run;

*Testar modellera med tredjegradspolynom;
proc mixed data=centrerad_data covtest;
    class id;
    model m = Ca Ca*Ca  Ca*Ca*Ca/ solution ddfm=bw outp=pred;
    random intercept / subject=id type=un gcorr;
run;
```

3.6.3 Deluppgift 1c

```
*Lägger till race;
proc mixed data=centrerad_data covtest ;
    class id race;
    model m = Ca Ca*Ca race / solution ddfm=bw;
    random intercept / subject=id type=un gcorr;
run;

*Lägger till sex;
proc mixed data=centrerad_data covtest ;
    class id race sex;
    model m = Ca Ca*Ca race sex/ solution ddfm=bw;
    random intercept / subject=id type=un gcorr;
run;

*Lägger till mammans ålder;
proc mixed data=centrerad_data covtest ;
    class id race;
    model m = Ca Ca*Ca race CAge_mom/ solution ddfm=bw;
    random intercept / subject=id type=un gcorr;
run;

*Lägger till birthyear;
proc mixed data=centrerad_data covtest ;
    class id race;
    model m = Ca Ca*Ca race Cbirthyear/ solution ddfm=bw;
    random intercept / subject=id type=un gcorr;
run;
```

3.6.4 Deluppgift 1d

```
*PROVAR NU MODELLERA MED OLIKA KOVARIANSMATRISSTRUKTURER FR R;

*CS;
proc mixed data=centrerad_data method=reml covtest;
    class id time race;
    model m = Ca Ca*Ca race / solution ddfm=bw;
    random intercept / subject=id;
    repeated time / subject=id type=cs rcorr;
run;

*UN;
proc mixed data=centrerad_data method=reml covtest;
    class id time race;
    model m = Ca Ca*Ca race / solution ddfm=bw;
    random intercept / subject=id;
    repeated time / subject=id type=un rcorr;
run;

*AR;

proc mixed data=centrerad_data method=reml covtest;
    class id time race;
    model m = Ca Ca*Ca race / solution ddfm=bw;
    random intercept / subject=id;
    repeated time / subject=id type=ar(1) rcorr;
run;


*SP POW;
proc mixed data=centrerad_data covtest;
    class id time race;
    model m = Ca Ca*Ca race / solution ddfm=bw;
    random intercept / subject=id;
    repeated time / subject=id type=sp(pow)(time) rcorr;
run;


*Toeplitz;
proc mixed data=centrerad_data method=reml covtest;
    class id time race;
    model m = Ca Ca*Ca race / solution ddfm=bw;
    random intercept / subject=id;
    repeated time / subject=id type=toep rcorr;
run;

```

3.6.5 Deluppgift 1e

```
*RESIDUALANALYS- kod från Jonas Bjermo;
proc mixed data=centrerad_data covtest plots=studentpanel;
    class id time race;
    model m = Ca Ca*Ca race / solution ddfm=bw;
    random intercept / subject=id;
    repeated time / subject=id type=ar(1) rcorr;
run;

```

4 Uppgift 2

4.1 Deluppgift 2a

Skatta en ”latent growth curve model” med ett latent intercept och en linjär latent lutning. Beskriv skattningarna och hur väl modellen är anpassad. (2p)

I denna deluppgift skattas en LGC (Latent Growrth Curve) modell. För modellen skattas ett latent intercept, en latent lutning (linjär) för de fyra tidpunkterna som mätts av barns utveckling i matematik. Modellformuleringen blir därför:

$$m_1 = \eta_{0i} + 0 \cdot \eta_{1i} + \epsilon_{1i}, \tag{12}$$
$$m_2 = \eta_{0i} + 1 \cdot \eta_{1i} + \epsilon_{2i}, \tag{13}$$
$$m_3 = \eta_{0i} + 2 \cdot \eta_{1i} + \epsilon_{3i}, \tag{14}$$
$$m_4 = \eta_{0i} + 3 \cdot \eta_{1i} + \epsilon_{4i}. \tag{15}$$

Ekvationerna (12-15) visar modellen som skattades i SAS. Här är m_1 till m_4 mätningarna för varje barn i vid de fyra tidpunkterna. η_{0i} är det latenta interceptet, alltså var barnens prestation i matematik börjar vid den första tidpunkten. I modellen är det inte uppenbart, men de latenta intercepten har en lastning på 1 (tänk att framför intercepten står det $\cdot(1)$. Detta innebär att intercepten påverkar alla tidpunkter lika mycket. η_{1i} är den latenta lutningen, vilket är hur barnets utveckling i matematik över tiden. Notera att varje modells latenta lutning multipliceras med antingen $0, 1, 2, 3$. Detta är alltså en linjär utveckling. I den första modellen m_1 finns ingen latent lutning på grund av att den multipliceras med noll, man kan se detta som att man identiferar interceptet vid det första mättillfället. De senare lastningarna 1,2 och 3 säger hur mycket värdet ökar för varje tidsenhet jämfört med startpunkten. Residualerna är ϵ_i.

4.1.1 Parameterskattningar

I denna del av uppgiften kommer jag att presentera modellens parameterskattningar i matrisen, därför kommer modellen nu presenteras i matrisform i tabell 22. I tabellen kan man se hur de olika delarna formuleras. Fetstil innebär att det är en matris.

Komponent	**Matris**
Observerade variabler	$\mathbf{Y}_i = \begin{bmatrix} m_{1i} \\ m_{2i} \\ m_{3i} \\ m_{4i} \end{bmatrix}$
Latenta faktorer	$\boldsymbol{\eta}_i = \begin{bmatrix} \eta_{0i} \\ \eta_{1i} \end{bmatrix}$
Faktorlastningar	$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$
Residualer	$\boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \\ \epsilon_{4i} \end{bmatrix}$
Residualkovarians	$\boldsymbol{\Theta}_\epsilon = \begin{bmatrix} \theta_\epsilon & 0 & 0 & 0 \\ 0 & \theta_\epsilon & 0 & 0 \\ 0 & 0 & \theta_\epsilon & 0 \\ 0 & 0 & 0 & \theta_\epsilon \end{bmatrix}$
Faktorkovarianser	$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{bmatrix}$
Modellens formel	$\mathbf{Y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i$

Tabell 22: Beskrivning av komponenterna för den LGC modellen. Inklusive matriser där parameterskattningar presenteras.

I tabell 23 presenteras modellens χ^2 test vilket i nollhypotesen undersöker om modellen faktiskt kan reproducera den sanna kovariansmatrisen i populationen. Det kan noteras att statistikan är 48.67 med ett signifikant p-värde, vilket innebär att nollhypotesen kan förkastas. Modellen kan alltså inte modellera den sanna kovariansmatrisen i populationen. Detta mått bör dock tas med en nypa salt, på grund av att stora stickprov tvingar nollhypotesen att förkastas. Man bör därför istället undersöka CFI och RMSEA (kommer snart).

Statistiska	**Värde**
Chi-square	48.669
df	4
Pr > Chi-square	0.0001

Tabell 23: Chi2 statistika med p-värde.

I tabell 24 presenteras RMSEA (Root Mean Squared Error Approximation) vilket tar hänsyn till urvalsstorleken i datamaterialet. Statistikan formuleras som:

$$\text{RMSEA} = \sqrt{\frac{\chi^2 - df}{df(N-1)}}$$

RMSEA kan tolkas som att ett värde som är mindre än 0.01 är okej, men helst vill man ha RMSEA som är under 0.05. Man kan tolka detta som att ju lägre RMSEA är desto bättre är anpassningen. RMSEA är 0.1496 vilket är mycket högt, och intervallet är $[0.1137; 01886]$. RMSEA visar att modellen verkar inte kunna modellar datan väl.

Tabell 25 presenterar Bentlers comparative fit index (CFI) vilket är ett mått som visar hur mycket bättre

Statistiska	Värde
RMSEA Estimate	0.1496
90% CI Lower	0.1137
90% CI Upper	0.1886
Probability of Close Fit	0.0001

Tabell 24: RMSEA utvärderingsmått med konfidensintervall

modellen är jämfört med en nollmodell där variablerna antas vara okorrelerade. Statistikan formuleras som:

$$\text{CFI} = \frac{(\chi^2 - df)_{\text{noll}} - (\chi^2 - df)_{\text{Modell}}}{(\chi^2 - df)_{\text{noll}}}$$

En modell med ett CFI över 0.9 visar att anpassningen är okej, men ett värde på över 0.95 visar en mycket bra anpassning. CFI är 0.96 vilket innebär att denna modell är bättre än en nollmodell.

Fit Index	Värde
Bentler Comparative Fit Index (CFI)	0.9592

Tabell 25: Comparative Fit Index (CFI).

I tabell 26 presenteras kovariansen mellan det latenta interceptet och lutningen. I tabell 27 presenteras medelvärdena för de latenta faktorerna. Men jag vill presentera dessa i matrisform så se nedan.

Var1	Var2	Estimate	Std. Error	t-Värde
f_{int}	f_{slp}	-8.575	2.056	-4.170

Tabell 26: Kovariansen mellan den latenta faktorn och lutningen.

Latent faktor	Parameter	Estimate	Std. Error	t-Värde
f_{int}	Medelvärde intercept	26.051	0.291	89.456
f_{slp}	Medelvärde slope	12.970	0.126	102.810

Tabell 27: Medelvärden för de latenta faktorerna.

4.1.2 Parameterskattning i matrisform

I tabell 28 presenteras kovariansmatrisen för LGC modellen. Variansen för det latenta interceptet (ψ_{11}) är 95.26 vilket visar att det finns en stor variation i barns startnivå av deras prestation i matematik, alltså att alla barn inte har samma startnivå i deras prestation i matematik. Variansen för den latenta ltuningen (ψ_{22}) är 6.37 vilket visar att det finns skillnader i hur snabbt barnen utvecklas i matematik. För att förstå varför detta är; om man tar roten ur får man en standardavvikelse på $\sqrt{6.37} = \pm 2.52$. Tidigare visades att barn har olika startnivåer i matematik. Medellutningen (visas snart) är drygt 13. Detta innebär att vissa barns prestation kanske ökar till 15.52, men vissa minskar till 10.48. Kovariansen (ψ_{12}) är -8.85 visar att barn vars startnivå är hög utvecklas långsammare, medan barn vars startnivå är lägre ökar snabbare.

$$\Psi = \begin{bmatrix} 95.62299 & -8.57517 \\ -8.57517 & 6.37273 \end{bmatrix}$$

Tabell 28: Kovariansmatrisen för de latenta faktorerna i LGC-modellen.

I tabell 29 presenteras medelvärdena för de latenta faktorerna. Interceptets medelvärde är 26.05 vilket innebär att ett barns genomsanittliga nivå i matematik börjar vid cirka 26 poäng. Medellutningen är

12.97 vilket innebär att ett barn i genomsnitt ökar med 13 enheter för varje tidsenhet. Med det sagt visar då modellen att barn faktiskt utvecklas över tid.

$$\mu_\eta = \begin{bmatrix} 26.05104 \\ 12.96970 \end{bmatrix}$$

Tabell 29: Medelvektorn för de latenta faktorerna i LGC-modellen.

I tabell 30 presenteras residualkovariansmatrisen. Det kan noteras att residualvariansen är densamma för alla tidpunkter, 35.58. Detta innebär att vid varje mätning (m_i) (enskilt) lyckas inte de latenta faktorerna fånga 36 enheters variation. Detta är en relativt hög siffra.

$$\Theta_\varepsilon = \begin{bmatrix} 35.58150 & 0 & 0 & 0 \\ 0 & 35.58150 & 0 & 0 \\ 0 & 0 & 35.58150 & 0 \\ 0 & 0 & 0 & 35.58150 \end{bmatrix}$$

Tabell 30: Residualkovariansmatrisen Θ_ε för LGC-modellen (lika residualvarians vid alla tidpunkter).

4.1.3 Slutsats

Modellen verkar inte vara särskilt lämplig. Först kunde nollhypotesen förkastas att modellen inte lyckas modellera den sanna kovariansmatrisen, men eftersom stickprovsstorleken kan påverka denna statistika undersökdes RMSEA. RMSEA visade även att modellen inte lyckas modellera datan väl. Det vill säga, modellen är inte lämplig.

4.2 Deluppgift 2b

Lägg till en kvadratisk faktor till modellen i uppgift 2 a). Beskriv skattningarna och hur väl modellen är anpassad. Jämför med resultaten i 2 a) och tolka skattningarna. (3p)

I denna deluppgift läggs nu en kvadratisk faktor till i modellen. Modellekvationen ändras där av och ser ut som:

$$m_1 = \eta_{0i} + 0 \cdot \eta_{1i} + 0 \cdot \eta_{2i} + \epsilon_{1i}, \tag{16}$$
$$m_2 = \eta_{0i} + 1 \cdot \eta_{1i} + 1^2 \cdot \eta_{2i} + \epsilon_{2i}, \tag{17}$$
$$m_3 = \eta_{0i} + 2 \cdot \eta_{1i} + 2^2 \cdot \eta_{2i} + \epsilon_{3i}, \tag{18}$$
$$m_4 = \eta_{0i} + 3 \cdot \eta_{1i} + 3^2 \cdot \eta_{2i} + \epsilon_{4i}. \tag{19}$$

De latenta faktorerna som presenteras i ekvation 16-19, tolkas som tidigare i deluppgift a) men nu har en latent kvadratisk faktor lagts till. Det vill säga, att modellen försöker fånga det icke-linjärt mönstret i datan.

I denna uppgift skattades modellen återigen i SAS. I tabell 31-35 presenteras parameterskattningar för modellen. Man kan notera i tabell 31 att variansen för både den latenta lutningen och den kvadratiska faktorn är negativ. Problemet i detta är att en varians är enligt definiton ≥ 0. En varians ska inte kunna vara negativ, men i detta fallen är den.

Vidare kan det notera i tabellerna 32 att p-värden för χ^2 inte kunnat beräknas. Detta beror på frihetsgraderna är noll. I tabell 33 kunde inte RMSEA beräknas, och i tabell 34 kunde inte CFI beräknas. Det kan alltså konstanteras att något är fel med modellen.

Parameterskattningarna för modellen bör därför inte tolkas grund av att modellen är inte korrekt. Jämfört man med modellen i deluppgift a) med denna modell kan man konstantera att, även om den linjära modellen inte var perfekt så är den mer lämplig än en modell med en kvadratisk faktor. Det vill säga, en mer komplex modell lyckas inte modellera det icke-linjära mönstret i datan.

Variable	Estimate	Std. Error	t Värde
Intercept variance (Var(f_{int}))	81.56138	7.68211	10.61706
Slope variance (Var(f_{slp}))	-2.01597	7.92506	-0.25438
Quadratic variance (Var(f_{quad}))	-0.47832	0.79671	-0.60037
Residual variance (e_1–e_4)	36.53814	2.31319	15.79557

Tabell 31: Parameterskattningar presenterade i utskriften i SAS för den kvadratiska modellen.

Statistiska	Värde
Chi-Square	27.1152
Degrees of Freedom	0
Pr > Chi-Square	—

Tabell 32: Chi2 test med p-värde.

Measure	Värde
RMSEA Estimate	—
90% CI Lower	—
90% CI Upper	—
Probability of Close Fit	—

Tabell 33: RMSEA utvärderingsmått.

Statistiska	Värde
Bentler Comparative Fit Index (CFI)	0.9753

Tabell 34: Bentlers Comparative Fit Index utvärderingsmått.

Var1	Var2	Estimate	Std. Error	t Värde
f_{int}	f_{slp}	13.16599	5.24985	2.50788
f_{int}	f_{quad}	-6.93514	1.53659	-4.51332
f_{slp}	f_{quad}	2.08371	2.41092	0.86428

Tabell 35: Modellens kovarianser mellan de latenta faktorerna.

Parameter	Estimate	Std. Error	t Värde
Mean of intercept factor ($\mu_{f_{\text{int}}}$)	23.95115	0.28706	83.43688
Mean of slope factor ($\mu_{f_{\text{slp}}}$)	13.47314	0.14279	94.35422
Mean of quadratic factor ($\mu_{f_{\text{quad}}}$)	1.82923	0.05739	31.87329

Tabell 36: Medelvärden för de latenta faktorerna.

4.3 Deluppgift 2c

Försök lägga till en eller flera bakgrundsvariabler från modellen i uppgift 2 b) för att förbättra modellens anpassning. Beskriv hur du har gått till väga samt vilka resultat du får. Motivera dina val. (Tips: se Paper 452-2013). (3p)

Det visades tidigare, i deluppgift b) att den modell med en kvadratisk latent faktor inte lyckas modellera det icke-linjära mönstret i datan. Modellen är inte bara olämplig utan fel. Detta kunde man se i samband med att variansen för lutningen och den kvadratiska faktorn var negativ. Nu testas om modellen förbättras ifall man lägger till en förklarande variabel.

4.3.1 Lägger till sex som förklarande variabel

Man kan tänka sig att pojkar kanske presterar sämre än flickor, därför kan det vara av intresse att lägga till kön i modellen.

I tabell 37-40 presenteras parameterskattningar och utvärderingsmått för den nya modellen. I tabell 37 visas variansskattningarna för modellen. Man kan konstantera att problemet med negativ varians kvarstår för lutningen och den kvadratiska faktorn. Tabellerna 38 och 39, är därför inte av intresse att tolka. I tabell 40 ser man att uvärderingsmåtten nu har tillhörande p-värden. χ^2 statistikan visar att nollhypotesen kan förkastas, men som tidigare nämndes, kan denna förkastas av misstag om datamaterialet är stort. Istället bör man undersöka RMSEA och CFI. För RMSEA är mycket bra vilket visar att modellen inte är tillräckligt bra. CFI ser bra ut, men detta beror mest troligtvis på grund av den låga frihetsgraden (1).

Man kan dra slutsatsen att en modell med en kvadratisk latent faktor inte förbättras för att man lägger till en variabel till den. Problemet med den negativa variansen kvarstår vilket tyder på att en mindre komplex modell är mer lämplig, alternativ en annan metod såsom GLMM.

Parameter	Estimate	Std. Error	t Värde
Var(e1–e4)	36.5381	2.3132	15.796
Var(f_int) (e5)	80.8984	7.6419	10.586
Var(f_slp) (e6)	-2.0536	7.9234	-0.259
Var(f_quad) (e7)	-0.4799	0.7966	-0.602
Var(sex)	0.2504	0.0159	15.796

Tabell 37: Variansskattningar för latent intercept, lutning och kvadratisk faktor samt mätfel.

Var1	Var2	Estimate	Std. Error	t Värde
f_int	f_slp	13.0081	5.2395	2.483
f_int	f_quad	-6.9031	1.5328	-4.504
f_slp	f_quad	2.0913	2.4106	0.868

Tabell 38: Kovarianser mellan latenta faktorer.

Latent faktor	Parameter	Estimate	Std. Error	t Värde
f_int	sex	-1.6272	0.9619	-1.692
f_slp	sex	-0.3875	0.8367	-0.463
f_quad	sex	0.0786	0.2632	0.299

Tabell 39: Parameterskattningar av effekten av kön i modellen.

Fitstatistik	Värde
Chi-Square	27.7395
DF	1
p-värde	0.0001
RMSEA	0.2315
SRMR	0.0352
CFI	0.9756
GFI	0.9975
AGFI	0.9507
AIC	65.7395
BIC	145.8170

Tabell 40: Fit-statistik för den kvadratiska LGC-modellen med kön som förklarande variabel.

4.4 Deluppgift 2d

Du har nu skattat ungefär samma modeller med två olika metoder. Jämför resultaten. Beskriv eventuella för och nackdelar med de både metoderna för datasetet som använts. (2p)

I den första uppgiften skattades en två-nivå mixed modell med ett slumpmässigt intercept. Modellen var ett grad tvåpolynom vilket lyckades fånga det icke-linjära mönstret i datan. För kovariansstrukturen användes en AR(1) struktur. Modellen gav mycket rimliga skattningar och visade att det barns prestation i matematik förbättras med deras ålder, där man även kunde se att ras också kunde skilja barn åt.

I den andra uppgiften skattades Latent Growth Curve modeller. Denna del av uppgiften visade att denna typ av metod inte lyckades modellera datan väl. Det visades även att LGC modellen inte lyckades modellera det icka-linjära mönstret i datan.

Man kan dra slutsatsen att för detta givna datamaterial är den mest lämpliga modellen en nivå-två mixed modell, vilket även visades vara mycket lämplig enligt residualanalysen. LGC modellen gav en dålig anpassning till detta datamaterial.

4.5 Bifogad SAS kod för uppgift 2

4.5.1 Deluppgift 2a

```
1
2  proc calis data=data method=ml;
3     lineqs
4         m1 = f_int + 0*f_slp + e1,
5         m2 = f_int + 1*f_slp + e2,
6         m3 = f_int + 2*f_slp + e3,
```

```
        m4 = f_int + 3*f_slp + e4;
   std
      f_int  = intercept_varians, /*jag ändrar namn för tydlighet*/
      f_slp  = slope_varians, /*jag ändrar namn för tydlighet*/
      e1-e4  = 4*var_e;
   mean
      f_int  = intercept_medel, /*Sätter lämpligt namn*/
      f_slp  = slope_medel; /*Sätter lämpligt namn*/
   cov
      f_int f_slp = kovarians_intercept_slope; /*Sätter lämpligt namn*/
run;
```

4.5.2 Deluppgift 2b

```
*Latenta faktorer med polynom;
proc calis data=data method=ml;
   lineqs
       m1 = f_int + 0*f_slp + 0*f_quad + e1,
       m2 = f_int + 1*f_slp + 1*f_quad + e2,
       m3 = f_int + 2*f_slp + 4*f_quad + e3,
       m4 = f_int + 3*f_slp + 9*f_quad + e4;
   std
      f_int  = intercept_varians,
      f_slp  = slope_varians,
      f_quad = quad_varians,
      e1-e4  = 4*var_e;
   mean
      f_int  = intercept_medel,
      f_slp  = slope_medel,
      f_quad = quad_medel;
   cov
      f_int  f_slp  = kov_int_slp,
      f_int  f_quad = kov_int_quad,
      f_slp  f_quad = kov_slp_quad;

run;
```

4.5.3 Deluppgift 2c

```

*UPPGIFT del c);
proc calis data=data method=ml;
   lineqs
       m1 = f_int + 0*f_slp + 0*f_quad + e1,
       m2 = f_int + 1*f_slp + 1*f_quad + e2,
       m3 = f_int + 2*f_slp + 4*f_quad + e3,
       m4 = f_int + 3*f_slp + 9*f_quad + e4,
       f_int  = b0_int*Intercept + b1_int*sex + e5,
       f_slp  = b0_slp*Intercept + b1_slp*sex + e6,
       f_quad = b0_quad*Intercept + b1_quad*sex + e7;
   std
      e1-e4   = 4*var_e,
      e5-e7,
      sex;
   mean
      sex;
   cov
```

```
19          e5 e6,
20          e5 e7,
21          e6 e7;
22
23  run;
```