Matematisk statistik
Stockholms universitet

APC modeller för Hodgkins Lymphom i Sverige 1957-2001

Filip Cederquist

Examensarbete 2004:13

Postadress:

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm
Sverige

Internet:

http://www.math.su.se/matstat

Matematisk statistik
Stockholms universitet
Examensarbete **2004:13**,
http://www.math.su.se/matstat

APC modeller för Hodgkins Lymphom i Sverige 1957-2001

Filip Cederquist*

juni 2004

Sammanfattning

Det finns i huvudsak två sätt att analysera data vad gäller ålder, period och kohorteffekter. Det första och mest precisa är anpassning av en parametrisk modell. Denna metod har dock nackdelar då presentationen av resultaten inte är helt enkel och tappar därför mycket information. Detta då variablerna av intresse är linjärt beroende tidsvariabler. Två av de största problemen som uppdagas är att effekterna inte är direkt identifierbara och att data uppvisar överspridning. Det andra sättet att analysera data är grafiskt. Graferna i sig är ganska enkla men problemen är desamma som för parametertolkningen. De bästa sätten att dra inferens om data är genom andra ordningens differenser bland parametrarna respektive tvådimensionella grafer. Det ständigt framträdande problemet med icke identifierbara effekter verkar till stor del komma av indelningen av data samt vilka effekter vi önskar veta något om.

Abstract

Information about age, period and cohort effects can be extracted from data mainly in two ways. The first and the perhaps most precise is the parametric approach. This, however, is proven to be cumbersome and not very informative when the effects we are interested in are highly (linearly) dependent time variables. Th two most evident problems that arise when fitting a model is how to present parameter estimates that suffer from non-identifiability and the problem of over-dispersion.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.E-post: fk02fce@math.su.se Handledare: Mikael Andersson.

The second angle is purely graphical. The idea is straightforward although the problems are the same as in the parametric approach. The most favorable way in which data can be analyzed is found to be in ways of second order differences in the parameters and two-dimensional plots, respectively. The problem of non-identifiability depicted time and again is mainly a result from the grouping of data but also from the choice of parameters and effects we consider interesting.

Innehåll sid.

1. Inledning

1.1. Problem

Problemet jag vill behandla i denna uppsats är tolkning av cancerdiagnosdata över tiden. Jag vill dels belysa svårigheterna i att ta till sig olika presentationer av data, dels praktiska problem vid statistisk modellanpassning och grafisk presentation. Frågor man kan tänkas formulera är om det exempelvis existerar en skiftning i risker att diagnosticeras med cancer mellan generationer eller om risken för diagnos helt enkelt har ökat/minskat över tiden. Den senare av dessa frågor är av stort intresse. Modellerna används här till cancerdiagnoser men kan självklart användas till mer generell data.

1.2. APC-modeller

Denna uppsats kommer att behandla så kallade APC-modeller. APC står för age-period-cohort dvs. ålder, period och kohort. De två första elementen är självförklarande medan det tredje i detta fall står för en generationsindelning.
Problemet jag vill beskriva med denna typ av modell är cancerdiagnoser och dess utveckling över tiden. Tiden representeras av tre olika variabler som i mycket glider in i varandra;

Ålder är naturligtvis den i år mätta åldern som en individ diagnostiseras med cancer.

Period är det datum som diagnosen ställs.

Kohorttillhörigheten betecknas här av födelseår.

Som vi ser kan dessa variabler parvis bestämma den tredje. Vet vi dagens datum och när en person är född kan vi avgöra hur gammal denna är. Liknande samband gäller givetvis mellan de övriga. Att de tre tidsvariabler vi valt att arbeta med i APC modeller är beroende av varandra vållar problem och fenomen vid alla typer av analyser av sådan data.
Vanligtvis när man talar om APC modeller är det frågan om att utröna förändringar som skett under en lång tid. För att jämna ut data delas denna in i lämpliga intervall. Då samtliga variabler är tidsvariabler uppträder framförallt ett fenomen. Det går exempelvis, teoretiskt, att åldras i modellen utan att tiden går. Fenomen som detta görs man inte uppmärksam på om data behandlas direkt i något programpaket t ex SAS. Ett statistikprogram kan omöjligen uppmärksamma sådana sublima tolkningar av data.
För att exemplifiera åldringsfenomenet tittar vi i tabell 2. Indelning av data är som beskrivs i tabellens not. Antag att vi befinner oss i åldersgruppen 90-94 och tillhöra kohorten 1862/63-70/71 under perioden 1957-61 (den vid första anblicken besynnerliga kohortnotationen kommer att förklaras i kommande avsnitt). Antag vidare att en individ är 94 år gammal, född 1863/64 och observeras 1958. År 1959 skulle denna person vara 95 år gammal och observeras under år 1960, allt av naturliga skäl. Fenomenet att personen faktiskt bytt åldersgrupp till den fetstilta utan att byta periodgrupp betyder inget annat än att vi med grupperingsförfarandet skapat illusionen av en minst sagt bisarr tidsaxel.
Frågan om det exempelvis går att visa en trend i perioderna kan bli svår att svara på då det på liknande sätt som beskrivits ovan går att byta period utan att åldras. Vi kan givetvis byta födelseår, kohort, med liknande resultat. Vi kan här ana att modellanpassningar och grafiska presentationer inte kommer att vara helt okomplicerade.

Skulle man vara ovetande om de fallgropar en APC analys innebär kanske man skulle börja med att titta på en graf över de olika variablerna. I figur 8 a-c ser vi att kanske borde vi vänta oss en signifikant skillnad mellan män och kvinnor. Vi ser också att periodtrenden med ögonmått mätt torde vara negativ. Man skall här komma ihåg att data graferna bygger på inte är viktade efter t ex skiftande åldersstruktur och dyrlikt. Fascinerande nog är det inte ett negativt samband vi finner vid modellanpassningar. Vad som uppträder när man försöker bena ut renodlade periodeffekter är snarare en positiv trend.

2. Dataindelning och grafisk presentation

2.1. Data

Sammanlagt fanns niotusen observationer tillgängliga. Insamlingen av data är utförd mellan 1957 och 2001. Data har varit tillgänglig på formen som presenteras i tabell 1.

SEX	YEAR	AGE	POP	LTYPE	CASES	PER-INT	AGE-INT	COH-INT	COUNTRY
Male	1957	0	55113	HL	0	1957-57	0-0	1956-57	Sweden
Fe-male	1957	0	52245	HL	0	1957-57	0-0	1956-57	Sweden
…	…	…	…	…	…	…	…	…	…

Tabell 1: Rådata

Variabler av särskilt intresse är AGE, YEAR och COHINT. Dessa tre är de så kallade APC –variablerna. APC –modellen och dess beståndsdelar kommer att presenteras i nästa avsnitt. I data uppträder bara en typ av lymphom nämligen Hodgkins lymphom (HL). Variabeln LTYPE tas alltså inte med i någon del av analysen.
Könsvariabeln SEX är givetvis intressant men vid könsindelning blir det mycket glest mellan fallen. Detta leder till att modellanpassningar blir mycket svåra. Jag har därför valt att slå ihop könsgrupperna till en när jag gjort den huvudsakliga analysen. I avsnitten om parametrar tar jag upp en modellanpassning där jag skattat en könsparameter i syfte att avgöra om det föreligger någon skillnad mellan män och kvinnor.

2.2. Indelning av data

Det enklaste sättet att ställa upp data är i kontingenstabellform. Perioder löper horisontellt och ålder vertikalt. Kohorter löper diagonalt från nedre vänstra hörnet till det övre högra. Givet kohort kommer en individ att röra sig med tiden längs en linje som startar i periodaxeln och löper snett nedåt höger (givet uppställning enligt tabell 2).
Exempelvis är en person som uppnått den mogna åldern 95-99 i perioden 1957-61 hörande till kohorten 1857/58-65/66. En 90-94 åring under samma period tillhör på samma sätt kohorten 1862/63-70/71. Delningstecknet för födelseår kommer av att en person född under det tidigare av åren uppnår den aktuella åldern under föregående period och bibehåller denna under en del av observerad period. T ex kan en 99 åring observerad i perioden 1957 fylla 99 under 1956. Den 99 åriga individen kommer då att vara 99 under en del av 1957 för att sedan fylla 100 under 1957. På samma sätt räknas alla individer som fyller 99 under året 1957 till åldersgruppen. De som fyller år i aktuell period är födda det senare av kohortåren. På samma sätt är de som träder in i perioden med den givna åldern födda under det tidigare av de två åren.

Period	**1957**	**1958**	**1959**	**1960**	**1961**	**1962**

Ålder						
93	1863/64	1864/65	1865/66	1866/67	1867/68	1868/69
94	1862/63	1863/64	1864/65	1865/66	1866/67	1867/68
95	**1861/62**	**1862/63**	**1863/64**	**1864/65**	**1865/66**	1866/67
96	**1860/61**	**1861/62**	**1862/63**	**1863/64**	**1864/65**	1865/66
97	**1859/60**	**1860/61**	**1861/62**	**1862/63**	**1863/64**	1864/65
98	**1858/59**	**1859/60**	**1860/61**	**1861/62**	**1862/63**	1863/64
99	**1857/58**	**1858/59**	**1859/60**	**1860/61**	**1861/62**	1862/63

Tabell 2: ålder, period och kohortindelning vid femåriga intervall. De fetstilta födelseåren tillhör en och samma grupp efter indelning. Tydligt i tabellen är att kohorter glider genom flera ålders- och periodgrupper.

Vi ser här att beroende på hur vi ställer upp data så kommer antingen ålder, period eller kohort att glida genom flera olika grupper. Vid en femårsindelning hör de fetstilta födelseåren till APC –gruppen [94-99,57-61,1857-66] där indelningen är [ålder, period, kohort].

Alternativt kan man skriva grupptillhörigheterna som [20,1,1]. Vid femårsindelning har vi alltså tjugo ålders, nio period och tjugoåtta kohortgrupper. I gruppen [20,1,1] ser vi att födelseåret 1865/66 dyker upp en gång. Samma födelseår representeras mer eller mindre i intilliggande grupper. Tabellutdraget är från ett hörn. Grupper som ligger i mitten av tabellen innehåller på liknande sätt kohorter som förekommer i ett stort antal andra ålder/period grupper. Detta ställer till med problem när vi vill uttala oss om t ex periodeffekter.
Jämförelserna mellan perioderna givet åldersgrupp kommer även att innehålla en oönskad kohorteffekt. Problemet blir uppenbarligen att effekterna inte går att skilja åt.

Valet att kohorterna i detta fall är glidande är främst beroende på att data oftast är sammanställt, eller snarare avdelat på ett sådant sätt att andra indelningar i stort är poänglösa. Oavsett vilken av de tre grupperna som är att betrakta som glidande kommer vi inte ifrån problemet med att effekterna också glider in i varandra.

Jag har valt att göra tre olika indelningar av data ett, tre och femårsintervall. Intervallängden syftar här till den i år mätta ålder och periodindelningen. Kohortindelningen följer automatiskt som intervall på två, sex respektive tio år. En indelning på ett år verkar vid första anblicken bäst. Detta då kohorterna inte alls glider genom flera ålder eller periodgrupper. Rent praktiskt är denna indelning väldigt klumpig att arbeta med. I den grafiska och parametriska presentationen som följer märks detta tydligt. I den tidigare formen av presentation är det svårt att uttala sig om data även med en grov indelning på fem år. I den senare kommer problemen av att årliga fenomen tillåts slå igenom. Detta leder till att frågor om trender under lång tid blir svåra att uttala sig om.

Frågan som man försöker svara på vid användande av APC-modeller är om det går att isolera utvecklingen av sjukdomen till ålder, period och kohorteffekter.
Svaret på detta är undvikande. Ja, det går att med matematiska verktyg skatta APC-effekter. Effekterna blir å andra sidan i all sin enkelhet mer eller mindre svårtydda i sin presentation. Detta beror främst på den logistiska regressionens egenheter [*ref. 3, 4*].
Värt att notera är att modellerna inte är till för att prediktera framtida sjukdomsfall. APC – modellernas resultat, i alla bemärkelser, är först och främst ämnade för att utreda hur trender under långa perioder utvecklat sig.

Oavsett indelning kommer vi att få en uppställning som på tabellform enligt tabell 3.

(ålder,period,kohort)	p =1	p = 2	…	p = P-1	p = P
a = 1	(1,1,A)	(1,2,A+1)	…	(1,P-1,A+P-2)	(1,P,A+P-1)
…	(2,1,A-1)	(2,2,A)	…	(2,P-1,A+P-3)	(2,P,A+P-2)
…	…	…	…	…	…
…	…	…	…	…	…
…	(A-1,1,2)	(A-1,2,3)	…	(A-1,P-1,P-1)	(A-1,P,P+1)
a = A	(A,1,1)	(A,2,2)	…	(A,P-1,P-1)	(A , P , P)

Tabell 3: Generell tabellindelning. Lägg märke till att kohorterna är en funktion av de två övriga tidsvariablerna

I tabellen ovan är A likställt med den äldsta åldersgruppen. P är på samma sätt den senaste perioden. Kohorterna följer av antal ålder och periodgrupper enligt k=A-a+p, där a och p är den i cellen aktuella ålder respektive periodgruppen. Den lägsta kohorten är givetvis ett och den högsta K=A+P-1. Med högsta kohorten menas den yngsta generationen.

2.3. Grafisk presentation

I ett försök att få en uppfattning om de data jag arbetat med har jag valt att först undersöka dem rent grafiskt.
Data presenteras i antal diagnoser samt hur stora delpopulationer dessa uppkommit ur.
Ett naturligt sätt att se på data är då

$$\hat{p}_{apc} = \frac{n_{apc}}{N_{apc}} \qquad a \in [1, A] \quad p \in [1, P] \quad c \in [1, C]$$

dvs. en empirisk sannolikhet att diagnostiseras med cancer. Sannolikheterna blir mycket små vilket har format ett modus operendi i litteratur rörande ämnet att multiplicera dessa med 100000. De sannolikheter som presenteras kan alltså läsas som antal fall per 100000 individer. Som i noten till tabell 3 bör poängteras att kohorterna är en funktion av ålder och period.

2.4. 3D-plottar

I figur 1 a-c ser vi att beroende på hur intervallen väljs för ålder och period antar planen en enklare och mer informativ form ju grövre indelning vi har.
Ettårsindelningen är mest rättvis men också svårast att uttala sig om rent grafiskt. De empiriska sannolikheterna i figur 1a verkar alla mycket små jämte ansamlingen av diagnoser bland höga åldersgrupper.
Redan vid treåriga intervall ser vi att planet antagit en lite mer lättillgänglig form.
Planet som är uppritat med femåriga intervall uppvisar en ännu mer behaglig form.
I figur 1a-c representerar X-axeln åldersindelning och Y-axeln periodindelning.
Kohorterna glider från undre högra hörnet till det övre vänstra.
Vi finner alltså den äldsta generationen längst ner till vänster och den yngsta högst upp till höger.

Oavsett vilken figur vi tittar på kan vi dra slutsatsen att personer som var 75-90 år under perioden 1957-65 löpte större risk att diagnostiseras med cancer än yngre under samma period. Vi kan i detta fall tala om en positiv ålderseffekt för individer som undersökts under de tidiga perioderna. Ålderseffekten är den omvända om vi istället tittar på individer som undersökts vid senare perioder.

Genom att följa 25-40 åringar (åldersgrupperna 6-8) i figur 1c ser vi att risken ökat under senare delen av 1900-talet. Detta tolkas som en positiv periodeffekt.
Kohorteffekter är i figur 1 förändringen i risk om man hoppar mellan diagonalerna som löper mellan nedre högra hörnet och det övre vänstra. Denna effekt blir i figur 1 a-c mycket svårfunnen. I figur 1 c finner vi den äldsta generationen, alltså kohorten född 1857/58-1865/66, i nedre vänstra hörnet. Den yngsta kohorten, 1992/93-2000/01 återfinns här i övre högra hörnet. Ett lättare sätt att orientera sig bland kohorterna är att plotta ålder mot kohort istället för ålder mot period. I figur 1 d presenteras ett ålder- kohortplan, som synes har höga samt mycket låga åldersgrupper en negativ kohorteffekt medan 20-29 åringar uppvisar positiv effekt.
Vi kan med vissa förbehåll säga att de mellersta och senaste generationerna, övre högra triangeln, löper mindre risk än de tidigare. Här talar vi om en negativ kohorteffekt.
Med denna grova analys kan man med utgångspunkt av data säga att;

1. Att vara äldre idag är mindre riskfyllt än förr.

2. Att vara yngre idag är mer riskfyllt än förr.

3. 50-talisterna löper mindre risk än sina föräldrar medan 60 och 70-talisterna uppvisar ökande risk.

Viktigt att poängtera är att de effekter som tas upp i den grafiska presentationen är ”ögonstatistik” och inga skattningar eller dyrlikt. De effekter som ses variera är de totala ålder- period kohorteffekterna, inte de isolerade som vi så gärna vill ha.
Grafisk presentation är mest ett sätt att koppla ett första grepp om data. Effekterna som dyker upp i grafisk presentation ger oss en fingervisning om vad vi har att vänta av resultaten från t ex en logistisk regression på data.

På sannolikhetsplanens form ser vi att anpassning av enkla modeller med linjära logodds kan bli svårt.
Tanken med att bredda intervallen är att jämna ut effekterna för att kunna uttala sig om hur de utvecklar sig över lång tid.
Bredare intervall för med sig att tolkningar av data blir trubbigare, varför modellanpassningar till data på denna form blir sämre.

2.5. 2D-plottar

Ett annat sätt att presentera samma data är i linjeplottar.
Jag har valt att presentera femårsindelningen då kortare intervall än fem år ger ett mycket rörigt intryck. Femårsdata kan vara svårt nog att ta till sig ändå. Tyvärr har jag inte lyckats konstruera några tjänliga plottar på hela data. Jag har valt att dela upp plottarna i tre delar. Figur 2 a-c samt figur 3 a-c kan alltså läggas i samma koordinatsystem men skulle då bli fullständigt obegripligt (se figur 2-3 d)

Figur 2 a-c visar risker betraktat som en periodeffekt. Grupperingen är åldersgrupper och löper horisontellt om ingen periodeffekt existerar. Som synes är ingen av åldersgrupperna att betrakta som parallell med periodaxeln. Möjligtvis kan åldersgrupperna 2, 6 och 7 betraktas som vågräta. Under femårsindelning motsvarar dessa åldrarna 5-9 samt 25-34. Tolkningen är att det i dessa grupper inte skiljer mycket mellan perioder, periodeffekten är alltså mycket liten. Generellt gäller detta mer eller mindre för alla personer mellan 0 och 34 år. I figur b och

c kan vi se en svag negativ trend för samtliga åldersgrupper. Individer i undersökningen över 34 år uppvisar alltså en minskad risk i dagsläget jämfört med samma ålder ett halvt sekel tillbaka i tiden.
Linjeplottarna är presenterade med logaritmerad skala på Y-axeln. Logaritmering för med sig att åldersgrupper med låga risker glider ifrån varandra i motsats till åldersgrupper med höga risker som å sin sida trycks ihop. Tolkningar rörande högriskgrupper, här de äldre grupperna, bör alltså tas med en nypa salt.
Data tillgänglig för analys uppvisar turligt nog en för stora sjok av åldersgrupper svag negativ periodeffekt. I figur 2 c ser vi tydligt hur skalan på Y-axeln trycker ihop riskerna. Tittar vi på figur 3 c ser vi en helt annan bild av periodeffekterna. I den senare av de två plottarna ritas även nollriskerna ut, dessa förekommer inte i figur 2 c då logaritmen av noll inte är definierad. De observerade nollriskerna förekommer bara i de två äldsta grupperna. Det går bara att spekulera kring orsakerna. Har klassningen av cancertyper eller kanske urvalet av äldre som undersöks ändrats?
Man skulle helt sonika kunna utesluta dessa nollobservationer och därigenom bara behålla åldersgrupperna 1-18 för att kringgå problemet. Jag har valt att behålla ”nollgrupperna”. Ett varningens finger bör dock höjas när dessa förekommer i grafer. Det verkar otroligt att diagnoser per 100000 individer helt plötsligt skulle falla från 16 till 0 under en tioårsperiod.

Figur 2-3 b uppvisar i likhet med figur 2-3 c en utjämning och en sammandragningseffekt av risker. Denna är kanske inte lika våldsam som för de äldre grupperna men dock närvarande. Det existerar som synes inga nollobservationer i de mellersta åldersgrupperna. I grafen med normal skala på Y-axeln märks periodeffekterna tydligare än om logaritmskala används.
För höga åldersgrupper, 11-13 d vs 50-64 åringar, kan man skönja en negativ trend i figur 3 a. Trenden bekräftas i figur 3 b som även framhåller en stor periodeffekt bland samtliga åldersgrupper.

I figur 2-3 a ser vi också logaritmskalans utjämnande effekter. I figur 2 a verkar det som om riskerna i den yngsta gruppen varierar mycket mellan perioderna medan de äldre grupperna uppvisar en flack bana. I figur 3 a är det istället de äldre grupperna som uppvisar tydliga periodeffekter. Man skall här komma ihåg att figur 3 a-c presenterar, trubbigt uttryckt, de egentliga riskerna dvs. utan skalförvrängning.

Ur linjeplottarna går det även att utläsa ålderseffekter. Givet en period kan man utläsa det lodräta avståndet mellan två åldersgrupper som förändringen i risk. Exempelvis är den absoluta skillnaden mellan åldersgrupp 4 och 18 enorm under period 1 (ca sex gånger större i grupp 18 jämfört med grupp 4) men inte alls lika avgrundsartad i period 9 (figur 3 a och c).

Skulle vi titta till kohortlinjeplottarna, figur 4-5 a-d, ser vi att dessa i stort är mer svårtydda än periodplottarna. Det går i likhet med perioderna att lägga dessa i samma koordinatsystem, frågan är om detta är särskilt belysande (figur 4-5 e). Den enda slutsats vi kan dra utifrån figur 4-5 e är att de yngsta åldersgrupperna (1-3 d vs 0-14 åringar) visar en generellt lägre risk än övriga. Skulle man vilja dela upp plottarna är detta inte lika enkelt som i periodfallet. Antingen plottar man , som i periodvarianten, vissa bestämda åldersgrupper i samma koordinatsystem (figur 4-5 aa-cc) eller så plottar man alla observationer i för ett bestämt antal kohorter (figur 4-5a-d). I det första fallet får man en god översikt vad gäller eventuella kohortförändringar. Problemet här är att man tappar svansarna från de intilliggande åldersgrupperna. Detta leder till att många ålderseffekter helt enkelt inte är med i plottarna. Vad gäller den andra indelningen kan även ålderseffekterna ses. Problemet med den senare

indelningen är att kohortskiftningarna blir mycket svåra att urskilja, särskilt för kohorter som innehåller alla nio åldersgrupper.
I min mening är figur 5 aa-cc mest informativ om man ser till kohorttrenderna. I stort sett alla de äldre åldersgrupperna visar en svängande men dock negativ utveckling ju senare de är födda, d vs ju högre kohort de tillhör. De yngsta grupperna (1-2 , 0-9 år) har en nästan horisontell utveckling medan 10-34 åringarna har en ryckig men i stort sett positiv trend. Man kan också notera att 15-24 åringar visar en mycket stor ökning i risk. Det ser alltså ut som att dessa grupper löper mycket högre risk ju senare de är födda.

Här ovan märker vi att alla tolkningar av data är relativa. Skattningarna jag har gjort i ett försök att anpassa en parametrisk modell lider av samma problem. Kraftfulla matematiskt statistiska modeller frälser oss alltså inte från att tolkningar och presentationer av desamma kan vara svåra.

3. Parametrisk presentation

I detta avsnitt ämnar jag presentera de analysverktyg jag använt mig av. Jag kommer gå igenom definitioner av modeller samt hur dessa ser ut vid anpassning till data. I detta avsnitt kommer även resultat från skattningar av parametrar att presenteras.

3.1. Logistisk regression

Valet av logistisk regression grundar sig på att svarsvariabeln i detta fall är binär. Med kategorisk/binär avses här en variabel som är ordinal. De ordinala värdena som antas är cancerdiagnos och ej cancerdiagnos. En positiv diagnos är kodad som 1 och avsaknaden av densamma är kodat som 0. Skälet till att klassningen 0 eller 1 betraktas som ordinal är då inget numeriskt avstånd kan observeras.
De förklarande variablerna ålder, period och kohort är däremot inte kategoriska utan är att betrakta som intervallvariabler. De är dock kodade som dummies och tas alltså inte upp som intervallvariabler. Möjligtvis kan svarsvariabeln ses som en nominal variabel. Val av klassificeringen påverkar dock inte valet av modell och lämnas därför.

3.1.1. Definition av (multipel) logistisk regression

För n(i) oberoende stokastiska variabler med

$$n_i \in Bin(N_i, p(x_i)) \qquad i \in [1,k]$$

$$\mathrm{x_i} = (1\ \mathrm{x_{1i}}\ \ldots\ \mathrm{x_{ri}})$$

där

$$\log\left(\frac{\mathrm{p(x_i)}}{1-\mathrm{p(x_i)}}\right) = b_0 + b_1\mathrm{x_{1i}} + \ldots + b_r\mathrm{x_{ri}}\ .$$

Skulle vi lösa ut sannolikheten ur ovanstående samband får vi följande.

$$p(x_i) = \frac{e^{b_0 + b_1 x_{1i} + \ldots + b_r x_{ri}}}{1 + e^{b_0 + b_1 x_{1i} + \ldots + b_r x_{ri}}}$$

Skattade sannolikheter och logodds har samma utseende som ovan men med skattade parametrar. I notationen innebär *r* antal parametrar. Hakparanteserna vi *i* syftar till heltalsdelen.

3.1.2. ML –Skattningar

Likelihooden för binomialfördelningen är som följer.

$$L(b) = \prod_{i=1}^{k} \binom{N_i}{n_i} p(x_i)^{n_i} (1 - p(x_i))^{N_i - n_i}$$

Vi söker ett maximum för denna vilket är ekvivalent med att söka maximum för den logaritmerade likelihooden (loglihooden).

$$l(b)=\sum_{i=1}^{k} n_i \log(p(x_i))+\sum_{i=1}^{k}(N_i-n_i)\log(1-p(x_i))+C=$$
$$=\sum_{i=1}^{k} n_i \log\left(\frac{p(x_i)}{1-p(x_i)}\right)+\sum_{i=1}^{k} N_i \log(1-p(x_i))=$$
$$=\sum_{i=1}^{k} n_i(b_0+b_1x_{1i}+...+b_rx_{ri})-\sum_{i=1}^{k} N_i \log(1+e^{b_0+b_1x_{1i}+...+b_rx_{ri}})$$

Observera att konstanttermen C inte skrivs ut efter andra likhetstecknet.
Derivering ger oss ett system av ekvationer på formen

$$\frac{\partial l(b)}{\partial b_j}=\sum_{i=1}^{k} n_i x_{ji}-\sum_{i=1}^{k} N_i p(x_i) x_{ji}$$

Detta ger oss ett maximum ty,

$$\frac{\partial^2 l(b)}{\partial b_j \partial b_l}=-\sum_{i=1}^{k} N_i x_{li} p(x_i) x_{ji}(1-p(x_i))\le 0$$

Om lösningar till förstaderivatorna satta till noll existerar är dessa unika.
Tolkningen av de skattade parametrarna är att betrakta som en förändring i odds. Te x om $e^{b_4}=1.2$ är tolkningen att oddset ökade med tjugo procent under period fyra jämfört med period 1. b_4 betyder parameter för logoddskvoten medan e^{b_4} syftar till parametern för oddskvoten.
För att tolkningen ska vara meningsfull måste vi vara bekanta med begreppet oddskvot.

$$OK(a,b)=\frac{p(x=a)/(1-p(x=a))}{p(x=b)/(1-p(x=b))}$$

Oddskvoten OK(a,b) skall betraktas som det relativa oddset att befinna sig i grupp a jämfört med grupp b. Definitionen av logistisk regression ger oss ett mer greppbart resultat av den logaritmerade oddskvoten.

$$\log\left(\frac{p(x_{ri}+1)/(1-p(x_{ri}+1))}{p(x_{ri})/(1-p(x_{ri}))}\right)=\log\left(p(x_{ri}+1)/(1-p(x_{ri}+1))\right)-\log\left(p(x_{ri})/(1-p(x_{ri}))\right)=$$

$$=b_0+b_r(x_{ri}+1)-b_0+b_rx_{ri}=b_r$$

Den logaritmerade oddskvoten är alltså ett annat uttryck för parametrarna i modellen. Jag kommer inte att vidare fördjupa mig i oddskvoter och relativa risker. En förklaring till varför

logoddset används är att sannolikheten i binomialfördelningen då begränsas till att existera mellan noll och ett. Detta är inte fallet vid exempelvis logaritmerad sannolikhet och ett linjärt parametersamband [*ref 12*]. Jag nöjer mig efter denna korta förklaring med att konstatera att oddskvoter är passande att använda i bland annat logistiska regressionsmodeller.

3.1.3. Logistisk regression i fallet AP-modell

För en modell som endast inkluderar ålder och periodeffekter ser den logistiska regressionsmodellen ut som följer.

$$n_{ap} \in Bin(N_{ap}, p(x_{ap})) \qquad a \in [1, A] \quad p \in [1, P]$$

$$\log\left(\frac{p(x_{ap})}{1-p(x_{ap})}\right) = m + a_a + b_p$$

Vid en anpassning till data skulle vi alltså utläsa de skattade riskerna att diagnosticeras med cancer som

$$p(x_{ap}) = \frac{e^{m+a_a+b_p}}{1+e^{m+a_a+b_p}}$$

som tabeller betraktat kan vi titta på skattningar av logoddset i tabell 4

Ålder / Period	1	…	P
1	$m+a_1x_{11}+b_1x_{11}$	…	$m+a_1x_{1P}+b_1x_{1P}$
…	…	…	…
A	$m+a_Ax_{A1}+b_1x_{A1}$	…	$m+a_Ax_{AP}+b_Px_{AP}$

Tabell 4: logodds. Vid skattningar läggs en felterm med väntevärde noll till . Observera att x i de celler de uppträder är dummyvariabler och därmed lika med ett.

Tabell 4-5 visar en generell bild av logodds/odds. En av cellerna fungerar som referenscell d vs att parametrarna i tabell 4 är satta till noll (interceptet bibehålls dock).
Logoddset är i sig kanske inte så intressant. Däremot är skattade sannolikheter av stort intresse.
Tabell 5 visar sannolikheter.

Ålder / Period	1	…	P
1	$\frac{e^{m+a_1x_{11}+b_1x_{11}}}{1+e^{m+a_1x_{11}+b_1x_{11}}}$	…	$\frac{e^{m+a_1x_{1P}+b_1x_{1P}}}{1+e^{m+a_1x_{1P}+b_1x_{1P}}}$
…	…	…	…
A	$\frac{e^{m+a_Ax_{A1}+b_1x_{A1}}}{1+e^{m+a_Ax_{A1}+b_1x_{A1}}}$	…	$\frac{e^{m+a_Ax_{AP}+b_Px_{AP}}}{1+e^{m+a_Ax_{AP}+b_Px_{AP}}}$

Tabell 5: Sannolikheter utbrutna ur definitionen av logodds.

3.1.4. Logistisk regression i fallet APC-modell

Definitionsmässigt är APC –fallet som vilken multipel regressionsmodell som helst

$$n_{apc} \in Bin(N_{apc}, p(x_{apc})) \quad a \in [1, A] \qquad p \in [1, P] \qquad c \in [1, C]$$

$$\log\left(\frac{p(x_{apc})}{1-p(x_{apc})}\right) = m + a_a + b_p + g_c$$

$$p(x_{apc}) = \frac{e^{m+a_a+b_p+g_c}}{1+e^{m+a_a+b_p+g_c}}$$

Skillnaden mellan APC och AP är ML-ekvationerna. I AP fallet är lösningarna till ML-ekvationerna unika vilket gör tolkningar av parametrar rättfram. För anpassning av APC -modellen gäller att lösningarna inte är unika. Lösningarna är unika då rangen för designmatrisen X är r+1 [*ref 8*] Med r menas som ovan nämnts antal parametrar exklusive intercept. Givetvis kommer åldersparametern få en undergrupp för varje åldersklass som uppträder i en vald indelning. Ålder representeras här av A-1 parametrar då en används som referenscell. För att illustrera detta tar jag hjälp av exemplet nedan.

Exempel:

Antag att vi har en uppdelning som ger oss två åldersgrupper och två periodgrupper. Betraktar vi detta som en tabell under generell definition av logistisk regression skulle *i* ha numreringen:

1	2
3	4

Indexering för AP och APC-modell skulle vara (a,p):

(1,1)	(1,2)
(2,1)	(2,2)

Inför ålderstabell med angivna åldersgrupper i cellerna enligt

$$\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$$

Inför även en periodtabell med utseende

$$\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$$

Kodar vi designvariabler 1 för förekomst av en parameter och 0 för avsaknad ger detta oss en designmatris av storlek k*r.. Antalet celler i en tabelluppställning är k.

En designmatris för en AP -modell på detta underlag skulle ha utseendet.

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

Här gäller att r=2 d vs att vi har två parametrar, ålder och period i vår modell. Dessa representeras av a_1 respektive b_2. Indexeringen av parametrarna avser till vilken period parametern hör. Som referens gäller att $a_2 = b_1 = 0$. Designmatrisen ovan har rangen 3 d vs r+1. Sista raden kan återskapas med de tre övriga. Däremot kan inte någon av kolumnerna återskapas med de övriga två. Notera att åldersgrupp 2 och period 1 är kodade som referenscell.
Skulle vi införa kohorter skulle dessa indexeras enligt följande.

$$\begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}$$

Designmatrisen skulle med kohorter ta formen nedan.

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Kodningen av dummies för en förklarande variabel med j nivåer ger j-1 dummies. För kohorterna här gäller indelning enligt tabell 6.

	x_{3i}=	x_{4i} =
Kohort 1	0	0
Kohort 2	1	0
Kohort 3	0	1

Tabell 6: Kohortdummykod till exemplet

Kohort 1 är i denna kodform att betrakta som referenskohort. Med referens menas att skattningar av övriga kohorter är en jämförelse med denna. Valet av referenscell i vid mening avgörs med avseende på population i samtliga celler. Den cell med flest undersökta individer bör utgöra en god grund för jämförelser. Som logodds betraktat sätts parametern för kohort 1 till 0. Vill man betrakta referenscellen som en multiplikativ effekt är den 1 av exponentskäl. Interceptet, eller konstanttermen, kan tolkas som en form av grundrisk.
Antal parametrar är r=4, 1 ålder, 1 period och 2 kohorter. Rangen för designmatrisen borde följaktligen vara r+1=5. Tyvärr är detta inte sant. T ex är kolumn 5 en linjärkombination av de övriga, kolumn 4 minus kolumn 2 plus kolumn 3 ger oss en (0 –2 0 0)’ vektor. Detta är inte överraskande då sambandet mellan kohorter, ålder och perioder alltid är k=A-a+p.
I exemplet ovan med kohorter är kohort 1 kodad som referenscell, d vs samma cell som i AP –fallet. Designmatrisen har alltså inte full rang vilket leder till att lösningarna till ML – ekvationerna ej är unika [*ref 5, 9*]

3.1.5. Modellanpassning

Modellanpassningar har gjorts i SAS för 1, 3 och 5-årsintervall. I tabell 7 presenteras anpassningen med tillhörande frihetsgrader, diskrepans och kvoten avvikelse/frihetsgrader.

Intervallängd	1 år	3 år	5 år
Frihetsgrader (df)	587	338	126
Avvikelse (D)	614,1781	369,9830	163,2669
D/df	1,0463	1,0946	1,2958

Tabell 7; De tre valda intervallindelningarnas passformer.

Som synes blir passformen sämre i takt med att intervallängden ökar. Detta skall vägas mot att längre intervall ger en utjämnande effekt. En utjämning är precis vad som är tanken med APC modeller då, som nämnts ovan, utvecklingen under en längre tid är av intresse. För ett och treårsindelningarna är datamängden begränsad. Att utesluta data är ett nödvändigt ont då det för korta intervall förekommer för stor andel nollobservationer. Gränserna för data är helt enkelt manuellt framtagna med kriteriet att SAS –algoritmen för logistisk regression skall konvergera.
Datamängden är medvetet avdelad så att åldersgrupper går förlorade istället för att mista värdefulla perioder. Skulle en period med många nollobservationer tas bort skulle de jämte åldersgrupper starkt underrepresenterade perioderna minska ytterligare. Att en åldersgrupp plockas bort är inte lika illa då individerna i denna grupp i nästa period ändå kommer in i datamängden. De åldersgrupper som faktiskt tagits bort består av mycket unga samt de åldersstigna. Typiskt för de unga grupperna är stora mängder undersökta och få eller företrädesvis inga diagnoser. Samma gäller för de äldre grupperna men med olikheten att det i dessa inte existerar lika stora mängder undersökta.
För femårsindelningen har allt datamaterial kunnat användas. Frihetsgraderna under respektive modell kommer av funktionen A*(P-1)-(P-1)-(K-2).
Skattade risker för femårsindelning finns uppritade i figur 1e. De skattade riskerna påminner starkt om de empiriska. De faktiska avvikelserna från samma anpassning återfinns i figur 1f. Skulle vi presentera residualer kan vi titta på Pearsonresidualerna som är på följande form.

$$S_{apc} = \frac{n_{apc} - \hat{n}_{apc}}{\sqrt{\hat{n}_{apc}}}$$

Vi ser att lejonparten av felen finns vid höga åldersgrupper kombinerat med tidiga perioder. Det är kanske inte så konstigt att så stor del av felen förekommer i det nedre vänstra hörnet om man tittar på figur 1a-c. Konturen i detta hörn vad gäller empiriska risker är mycket taggig , figur1a, alternativt böljande, figur 1b-c. I figur 1g finner vi Pearsonresidualerna uppritade mot ålder, period och följdaktligen kohorter. Den slumpartade formen på planet tyder på att modellen troligen inte har några större systematiska fel.

3.2. Generaliserade linjära modeller, GLM

Ett annat sätt att anpassa modeller är med GLM. Anledningen till att dessa tas upp här är att GLM modeller har intressanta egenskaper vad gäller spridning av data. Det krävs dock en presentation av GLM. Presentationen av modellen kommer inte att vara vidare rigorös då GLM inte står i fokus i detta fall.

3.2.1. Definition av GLM

Antag att fördelningsfunktionen för någon stokastisk variabel y kan skrivas som.

$$f(y\,|q) = R(q)\exp\left\{\sum_{j=1}^{v} q_j(q)t_j(y)\right\}h(y)$$

Det ger att y tillhör familjen exponentiella dispersionsmodeller (EDM). Den övre gränsen i summeringen avgör hur många parametrar fördelningen har. I binomialfallet är v=1.
Vidare för GLM gäller att:

- $y_i \in f(y_i\,|q_i, f, w_i)$
- $E(y_i) \equiv m_i$
- $g(m_i) = x_i'b$ någon g(.)

Betavektorn är en okänd parametervektor av samma längd som antalet parametrar i modellen. Vektorn x är en vektor med kända prediktionsvariabler.
För att GLM skall kunna användas gäller vidare att fördelningenfunktionen för y kan skrivas på formen nedan.

$$\exp\left\{\frac{w_i}{f}[qy - r(q)]\right\}h(y, f, w)$$

Om en fördelning är EDM medför detta att den går stt skriva om på GLM –form och vice versa.
Anledningen till att GLM presenteras är f. f betraktas inte som en parameter i normal mening utan snarare som ett spridningsmått. En viktig del av GLM är g(.) som kallas länkfunktionen. Om $g(\bullet) = \text{Ý}^{1}(\bullet)$ sägs man använda kanonisk länk. Kanonisk länk vid logistisk regression i GLM är såkallad logit -länk. Pricken ovanför r betecknar förstaderivatan. Använder man sig av kumulantgenererande funktioner (logaritmerade momentgenererande funktioner) finner vi på ett smidigt sätt att variansen för y och länkfunktionen har ett samband.

$$Var(y) = \text{Ÿ}(q)\frac{f}{w}$$

Vi kan här skriva om andraderivatan av r till:

$$\text{Ÿ}(q) = \text{Ÿ}(\text{Ý}^{1}(m)) = V(m)$$

V(m) kallas variansfunktionen. I uttrycket för variansen kan vi se f som en korrektionsparameter för modellens varians. Skulle vi i och med en modellanpassning exempelvis finna att skattningen av spridningsparametern är större än ett talar vi om överspridning. Med andra ord gäller att om $f > 1$ har data större varians/spridning än vad en enparametrig modell kan förklara. Mer om detta i avsnittet som berör parameterskattningar.

3.2.2. Skattningar av parametrar

ML –skattningar av parametrar i GLM utgår givetvis från likelihoodfunktionen. Antag att vi har n st oberoende observationer av y.

$$L(b;f)=\prod_{i=1}^{n} f(y_i \mid q_i, f, w_i) = \left[\text{antag; } \exists g_*(\bullet) \text{ så att } q_i = \theta\left(g^{-1}(x_i'b)\right) \equiv g_*(x_i'b)\right] =$$

$$= \prod_{i=1}^{n} f(y_i \mid x_i'b, f, w_i) =$$

Logaritmering och derivering ger skattningar.

$$\ell(b;f)=\sum_{i=1}^{n} \log\left[f(y_i \mid x_i'b, f, w_i)\right] = \sum_{i=1}^{n} \frac{w_i}{f}\left[x_i'by_i - r(x_i'b)\right] + C$$

$$\frac{\partial \ell(b;f)}{\partial \ell b_j} = 0$$

Skattningar av delarna i modellen följer.

$$\hat{m}_i = g^{-1}\left(x_i'\hat{b}\right) \qquad g(\hat{m}_i) = x_i'\hat{b} \qquad \hat{q}_i = g_*\left(x_i'\hat{b}\right)$$

Skattning av spridningsparametern går att göra på flera sätt. Vi börjar med att definiera huvudkomponenterna i skattningarna, Pearsons statistika respektive Diskrepansen.

.

$$\mathrm{X}^2 = \sum_{i=1}^{n} w_i \frac{(y_i - \hat{m}_i)^2}{V(\hat{m}_i)}$$

$$D(\hat{m}) = 2\sum_{i=1}^{n} \hat{m}_i \log\left(\frac{\hat{m}_i}{\hat{m}_i^{(0)}}\right)$$

Summeringen till n blir i APC –fallet över alla celler/observationer. Med t ex femårsindelning gäller n=180. I definitionen av diskrepans är nollhypotesen den perfekta eller alternativt uttryckt den mättade modellen. Skattningarna av f i de olika fallen är:

$$\hat{f}_P = \frac{\mathrm{X}^2}{n-r}$$

$$\hat{f}_D = \frac{D(\hat{m})}{n-r}$$ r är i båda fallen antal frihetsgrader.

Hur ser då skattningarna ut vid olika anpassningar?

intervall	1	3	5
P och D	1,0229	1,0462	1,1383

Tabell 8; Skattning av spridningsparametern vid olika intervallängd.

Skattningarna är lika i båda fallen och i och med att intervallen växer, ökar också överspridningen. Överspridningen i femårsfallet är så stor att man bör överväga att anpassa en modell med större varians. I det aktuella fallet med APC –modeller ligger kanske negativ

binomialfördelning närmast till hands. Den praktiska tolkningen av denna fördelning om vi skulle anpassa den till data är kanske inte helt enkel. Jag har därför utelämnat anpassning och tolkning av den negativa binomialfördelningen i denna text. Visst är det möjligt att anpassa en modell med flera parametrar än binomialfördelningens enda. Frågan är om det är någon idé då tolkningarna av parameterskattningarna redan vid logistisk regression är svårtolkade.

3.3. Parametrar

Parametrarna i alla tre modeller ovan är som nämnts inte förändringar i och med gruppbyten i vanlig mening. Parameterskattningarna är inte unika och måste därför presenteras på ett manér som kan verka ickeintuitivt.
Skattningar av t ex periodeffekter innehåller delar av ålder och kohorteffekter. För att komma ifrån dessa är det smidigt att presentera parametrar som relativa risker [*ref. 4*].

$$\frac{\left(e^{b_{p+1}} / e^{b_p}\right)}{\left(e^{b_p} / e^{b_{p-1}}\right)}$$

Även ålder och kohorteffekter lider av dessa problem. De presenteras på liknande sätt.

$$\frac{\left(e^{a_{a+1}} / e^{a_a}\right)}{\left(e^{a_a} / e^{a_{a-1}}\right)} \quad \text{respektive} \quad \frac{\left(e^{g_{c+1}} / e^{g_c}\right)}{\left(e^{g_c} / e^{g_{c-1}}\right)}$$

Skattningar av parametrar i femårsintervall presenteras i tabell 9. De effekter som avses ovan är de multiplikativa och alltså inte de additiva gällande de logaritmerade oddskvoterna. Skulle vi använda de additiva skulle vi som i fallet med multiplikativa effekter få ett uttryck som också tolkas som kurvatur. Kvoterna mellan de relativa riskerna kan alltså ses som acceleration kring gruppen a, p eller c. Kvoterna ovan presenteras för femårsindelningen i figur 6a-c. Det är tydligt att trender kan vara svåra att urskilja för samtliga effekter. Tittar vi på figur 6a kan vi med lite god vilja påstå att ålderseffekterna verkar uppvisa en puckelform. Puckeln tolkas som att man i de mittersta åldersgrupperna löper störst risk att diagnosticeras med cancer.
Figur 6b är ett försök att utreda periodeffekter. Periodriskerna är taggiga till utseendet vilket gör dem svårtolkade. Möjligtvis kan vi se att periodeffekterna har planat ut under senare delen av 1900-talet.

Åldersgrupp	*	**Kohortgrupp**	*
1	0,054	*1*	0,0002
2	0,2	*2*	28,51
3	0,41	*3*	60,34
4	1,10	*4*	34,18
5	1,5	*5*	33,22
6	1,28	*6*	22,04
7	1	*7*	17,29
8	0,79	*8*	11,81
9	0,63	*9*	9,1
10	0,49	*10*	6,35
11	0,45	*11*	5,22
12	0,39	*12*	3,73
13	0,41	*13*	2,75
14	0,39	*14*	2,12
15	0,36	*15*	1,84
16	0,28	*16*	1,43
17	0,24	*17*	1,18
18	0,14	*18*	1
19	0,09	*19*	0,82
20	0,02	*20*	0,77
Periodgrupp	*	*21*	0,76
1	0,39	*22*	0,72
2	0,71	*23*	0,83
3	0,87	*24*	0,95
4	1,11	*25*	0,97
5	1	*26*	0,72
6	1,12	*27*	0,53
7	1,24	*28*	1
8	1,31		
9	1,36		

Tabell 9; Skattningar av APC parametrar för femårsintevall. Referenscell är a=7, p=5 och c=18. Till höger om varje grupp följer den multiplikativa effekten.

Vad gäller kohorteffekterna uppvisar de samma taggighet som perioderna. Det är svårt att uttala sig om kohorterna förutom att det finns två toppar, en tidigt och en sent. Den första toppen ses som risken mellan kohort 5 och 4 (1877/78-85/86 respektive 1872/73-80/81) jämfört med kohort 4 och 3 (1872/73-80/81 respektive 1867/68-75/76). Den senare gäller relativa risken mellan kohort 28 och 27 (1992/93-2000/01 respektive 1987/88-95/96) samt kohort 27 och 26 (1987/88-95/96 respektive 1982/83-90/91). Man bör hålla i minnet att skattningarna av dessa kohorteffekter bygger på väldigt få observationer. Skakigheten i skattningarna samt de få observationerna kommer av att dessa kohorter ligger i hörnen av en tabell. Exempelvis innehåller kohort 1 och 28 (egentligen 1 och K) endast en cell. Det är därför vanskligt att säga något om kohorter som ligger i periferin i tabeller.
Ett försök att utjämna effekterna är att använda sig av produkten av två intilliggande relativa risker. Dessa presenteras som linjegrafer i figur 7a-c.
I figur 7a uppträder en tydlig puckel för de mellersta åldersgrupperna. Vi ser också en tydlig topp för unga åldrar.

Tittar vi tillbaka på figur 1c kan vi även här se denna topp, dock inte lika tydligt. I figur 7b ser vi förvånande nog en tydlig positiv trend vad avser perioder, denna är omöjlig att skönja i figur 1c. Vi ser även i figur 7b att periodeffekterna verkar stabilisera sig över tiden. I dagsläget löper man alltså större risk för cancer än i mitten av 1900-talet om vi ser till perioder d vs risken verkar ha ökat i takt med att tiden gått.
Kohorteffekterna uppvisar i figur 7c ingen trend alls om vi bortser från de skarpa hoppen i början och i slutet. Hoppen i början och i slutet av linjegrafen med avseende på kohorter kommer förmodligen av det lilla dataunderlag skattningarna baserar sig på. Det går att lägga in restriktioner på extremkohorterna genom att sätta dessa till noll, som additiv effekt [*ref 7*]. Jag har valt att inte göra detta då jag faktiskt vill visa svårigheterna med modellanpassningar. Det visar sig att många kohorter knappt går att skatta. Extremkohorten 28 i femårsindelningen har visat sig vara mycket tvivelaktig. Skattningarna har skiftande kvalitet vilket gör det svårt att säga något om huruvida t ex kohorterna överhuvudtaget har något i modellen att göra. Funktionen ”type3” i SAS ger teststorheter som gäller för grupper överlag. Vid genomförande av sådana test visar det sig att samtliga grupper är relevanta. Man bör dock beakta att testet är lite svårt att använda sig av om det inte ses i sitt samband med skattningarna av de enskilda parametrarna.
Ett annat sätt att testa om effekterna är framträdande är genom D-subtraktion. Genom att jämföra intilliggande modellers avvikelse från den mättade modellen, diskrepansen, kan man skaffa sig en uppfattning om en grupp av variabler är signifikanta. Svårigheten är att av de modeller man jämför måste den med flest parametrar vara en delmängd av den enklare av de två. Skulle vi anpassa en modell med enbart ålderseffekter skulle denna kunna jämföras separat med modeller innehållande ålder och period samt ålder och kohorter. Jämförelserna skulle bli ganska trubbiga om vi ser till renodlade period eller kohorteffekter. En D-subtraktion mellan den fulla APC modellen och AP samt AC modellerna skulle testa kohorteffekterna respektive periodeffekterna. Det är viktigt att komma ihåg att AP och AC modellerna inte kan jämföras sinsemellan då ingen av dessa är delmängd av den andra. D-subtraktionen kan bara utröna huruvida period och kohorteffekter har något att bidra med i APC modellen. I tabell 10 presenteras nödvändiga data för D-subtraktion vad avser femårsintervall.

Modell Df D	A 160 1205	
Modell Df D	AP 152 722	AC 133 432
Modell Df D	APC 126 163	

Tabell 10; Data för D-subtraktion. Med modell avses de effekter som inkluderas i anpassning.

Vill vi testa om periodeffekter bör inkluderas i APC modellen subtraheras APC från AC. Vi får efter subtraktionen en chitvåfördelad testvariabel med värdet 269 och 7 frihetsgrader. Detta tyder på en närvarande periodeffekt. På samma sätt testas om kohorteffekten är relevant. AP minus APC ger 559 och 26 frihetsgrader. Detta säger oss att även kohorter har en given plats i den slutgiltiga APC modellen.

3.3.1. Könsstratifiering

En parameter som faktiskt kan tolkas på normalt sätt är en eventuell könsparameter. Jag har skattat denna i femårsindelningen. Resultatet presenteras i tabell 11.

df	D	D/df
305	452,6794	1,4842

Tabell 11; Passform för femårsindelning med könsdummy.

Passformen sjunker markant men av intresse är könsvariabeln för kvinnor jämte män.
Könsdummien skattas till –0,4193 vilket i multiplikativa mått är 0,6575. Ett 95-procentigt konfidensintervall visar sig vara [0.62, 0.69] för den multiplikativa effekten. P-värdet för att effekten skulle vara ett d vs att det inte spelar någon roll om en person är man eller kvinna är mindre än en tusendel. Att vara kvinna medför i detta fall att man löper ca 35 procents lägre risk att diagnosticeras med cancer om man jämför med män.
Anledningen till att passformen sjunker vid införande av en könsdummy kan vara den att det för låga åldrar knappt existerar några diagnoser bland kvinnor. Med låga åldrar avses grupp ett alltså 0-4 åringar. På samma sätt fast utan genusaspekten tunnas diagnoserna ut i den högsta åldersgruppen. I de mellersta grupperna genererar dock inte könsindelningen några luckor. Dessa nya nollobservationer i extrema ålder och periodgrupper torde förklara varför passformen sjunker så dramatiskt i och med en könsdummy.
I ett försök att enbart förklara könsskillnader har jag anpassat en logistisk modell till data som enbart sorterats efter kön. Den renodlade könsmodellen visar liknande resultat som vid införande av könsdummy. Kvinnor löper i den senare 31 procents lägre risk att diagnosticeras med cancer jämfört med män.

4. Slutsatser och Kommentarer

Att det är mycket svårt att uttala sig om APC –effekter blir uppenbart för envar som sätter sig ned och arbetar med data på denna form. Skulle jag väga analysverktygen grafisk och parametrisk presentation föredrar jag nog den tidigare då den är mer lättillgänglig. Den parametriska anpassningen kräver ganska avancerade presentationsformer vilket gör att den blir svår att ta till sig. Visst går det att komma runt problemet med identifierbarhet genom att införa restriktioner på variablerna [*ref 11*]. Detta är berättigat i många fall men i denna uppsats lite överdrivet då jag framförallt är ute efter att beskriva problemet. Om man läser tex Clayton och Schifflers artiklar [*ref 3, 4*] finner vi presentationer av data som är av den välartade sorten. Personligen har jag inte kunnat slå fast lika tydliga resultat, kanske då jag inte haft förmån att själv välja data.
Som nämnts ovan anser jag att den grafiska presentationen är att föredra över den parametriska. Den grafiska framställningen är kanske mest rättvis i sin tredimensionella uppläggning. Problemet när man sätter tredimensionella plottar på pränt så tappar de mycket av sin mening. När jag arbetat med att rita upp data har jag alltid kunnat vrida och vända på planen som presenteras i figur 1a-g. Detta går förlorat när jag här presenterar plottarna. Det är också svårt att orientera sig när man tittar på uppritade plan i stora detaljerade koordinatsystem. Orienteringsproblematiken gör att figurerna mest är att se som en översikt. Ett problem som inte framgår särskilt väl när man arbetar med programpaket är att det sällan presenteras med vilken metod graferna ritas. Problemet med uppritningsmetoden är tydlig om man tittar på figur 1e jämfört med övriga medlemmar av figur 1. Den egentliga svårigheten vid databearbetning är inte plottandet utan indelning av data i lämpliga intervall. För mig kan det tyckas att om indelningen ändå är gjord är det en smal sak att göra modellanpassningar jämfört med de cirka tvåtusen rader kod som plottarna kräver. I de flesta programpaket är dock plottarnas kod det minsta problemet. Därför anser jag att plottarna är att prioritera vid arbete med APC -modeller. Naturligtvis är inte parameterframställning utesluten men man bör komma ihåg den begränsade inferens man kan dra av skattningarna.
Det finns en mängd olika typer av grafiska framställningar att använda sig av. De flesta ger dock väldigt lite information om man ser till hur lång tid det tar att sätta sig in i vad som egentligen framställs [*ref 10*].
Ett stort problem vid anpassning av binomialfördelade data är att man antar att alla individer i en cell uppvisar samma risk att drabbas av cancer. Antagandet är starkt och inte helt sant. Det verkar orimligt att en kedjerökande reaktorrengörare löper samma cancerrisk som en ickerökande bonde bosatt i Norrlands inland. Skulle vi istället anta en Poissonfördelning med gammafördelad parameter skulle vi behöva anpassa en negativ binomialfördelning. Detta ger stöd för denna typ av anpassning men ingen intuitiv tolkning av den negativa binomialfördelningens extra parameter.
I datta fall är parameteranpassningen inte särskillt bra. Det är därför svårt att säga något kärnfullt om de skattade effekterna.
För gällande data vill jag därför rekomendera en rent grafisk analys.
Den grafiska framställningen är för gällande data helt enkelt minst förvirrande jämfört med den parametriska.

Referenser

1. Christensen, R. ”Log-Linear Models and Logistic Regression, 2nd ed”, Springer, 1997

2. Clayton, D & Hills, M. ”Statistical Models in Epedemiology”, Oxford university press, 2002

3. Clayton, D & Schifflers, E. ”Models for temporal variation in cancer rates I: Age-period and age-cohort models”, Statistics in Medicine, 6, 449-467 (1987)

4. Clayton, D & Schifflers, E. ”Models for temporal variation in cancer rates II: Age-period-cohort models”, Statistics in Medicine, 6, 469-281 (1987)

5. Holford, T. R. ”The estimation of age, period and cohort effects for vital rates” , Biometrics, 39, 311.324 (1983)

6. Lemeshow & Hosmer ”Applied Logistic Regression”, Wiley, 1989

7. Liu, S. Semenciw, R & Mao, Y. ”Rewiew article; Increasing incidence of non-Hodgkin’s lymphoma in Canada, 1970-1996: Age-period-cohort analysis ”, Hematol Onkol, 21, 57-66 (2003), online: 20 januari 2003 på Wiley InterScience (interscience.wiley.com). DOI:10.1002/hon.703

8. Ohlsson, E. ”Log-linjära modeller och Logistisk regression”, Kompendium SU, 2003

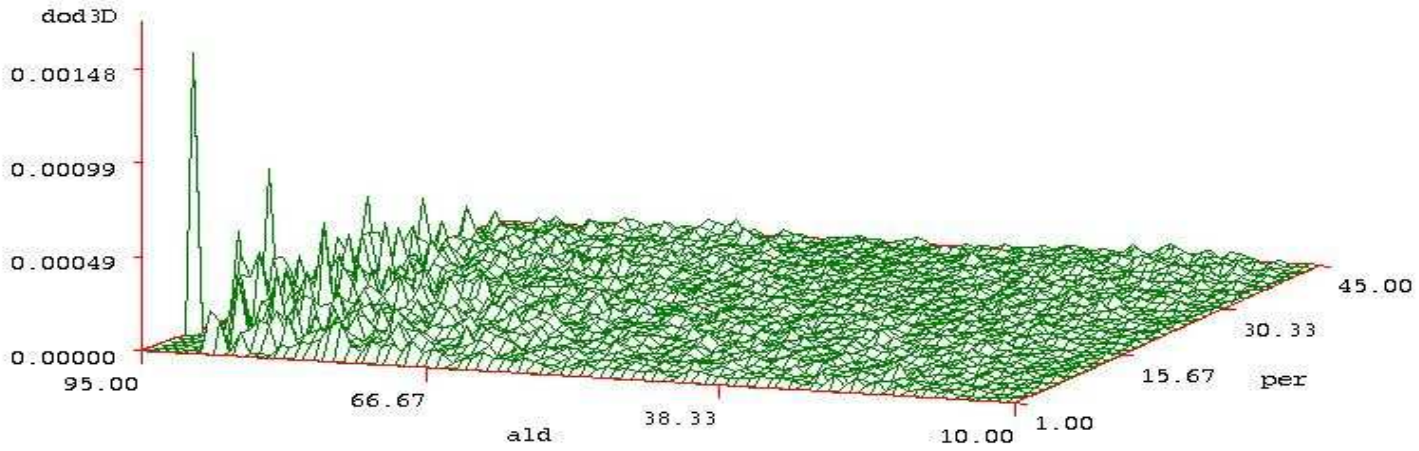
9. Osmond, C & Gardner, M. J. ”Age, period and cohort models applied to cancer mortality rates”, Statistics in Medicine, 1, 245-259 (1982)

10. Robertson, C & Boyle, P. ”Age-period.cohort models of chronoc diseaserates. II: Graphical approaches”, Statistics in Medicine, 17, 1325-1340 (1998)

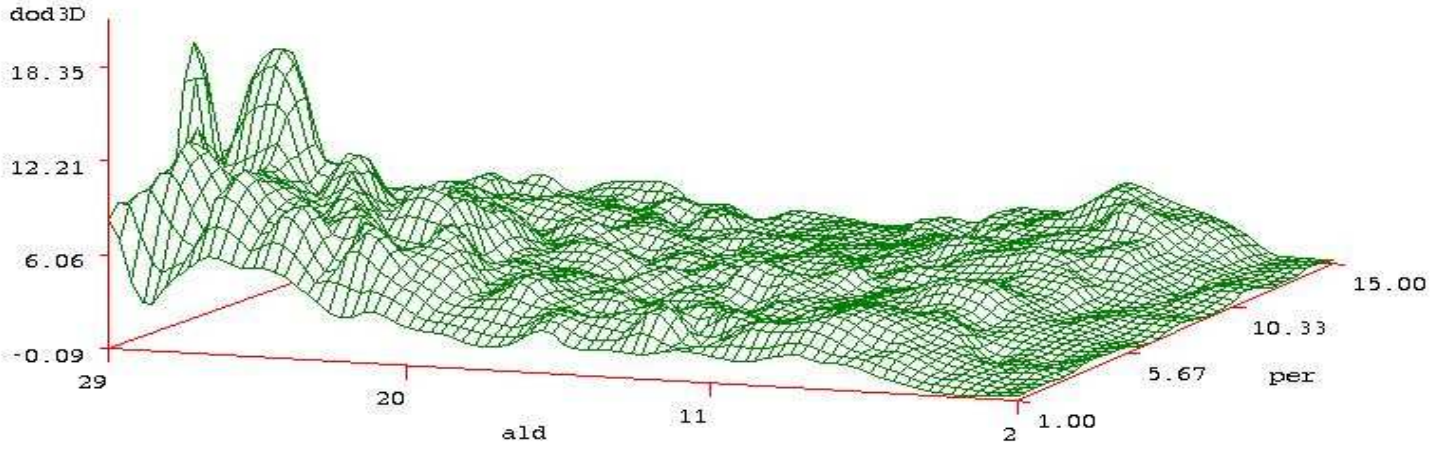
11. Rostgaard, K , Væth, M. , Holst, H. , Madsen, M. & Lynge, E. ”Age-period-cohort modelling of breast cancer incidence in the Nordic countries”, Statistics in Medicine, 20, 47-61 (2001)

12. Tamhane, A & Dunlop, D. ”Statistics and Data Analysis from Elementary to Intermediate”, Prentice Hall, 2000

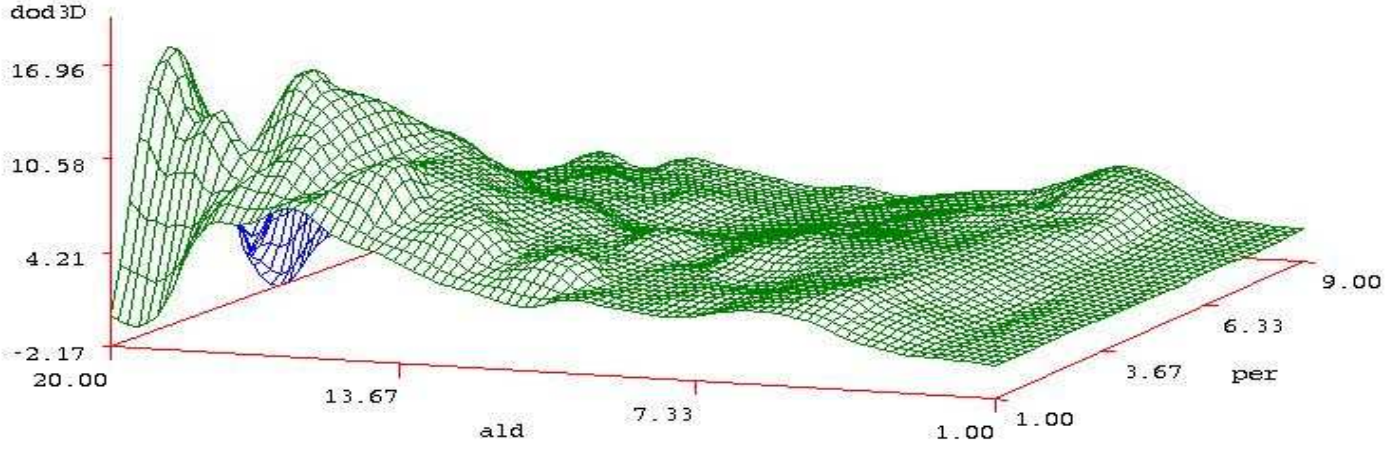
figur 1a



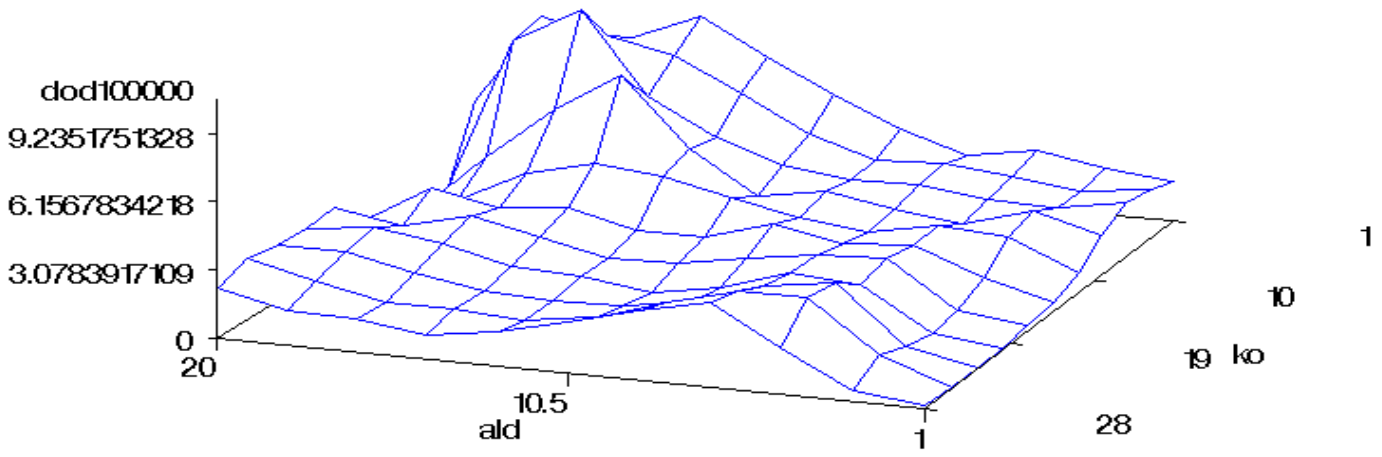
figur 1b



figur 1c

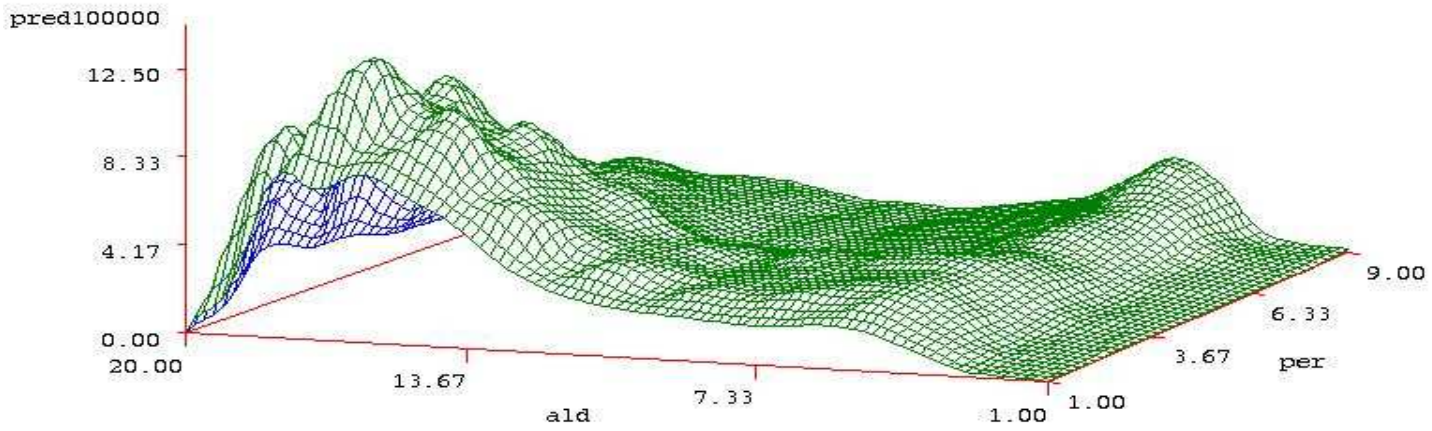


figur 1d
ald*kohort OBS:l & r corners do not exist in data



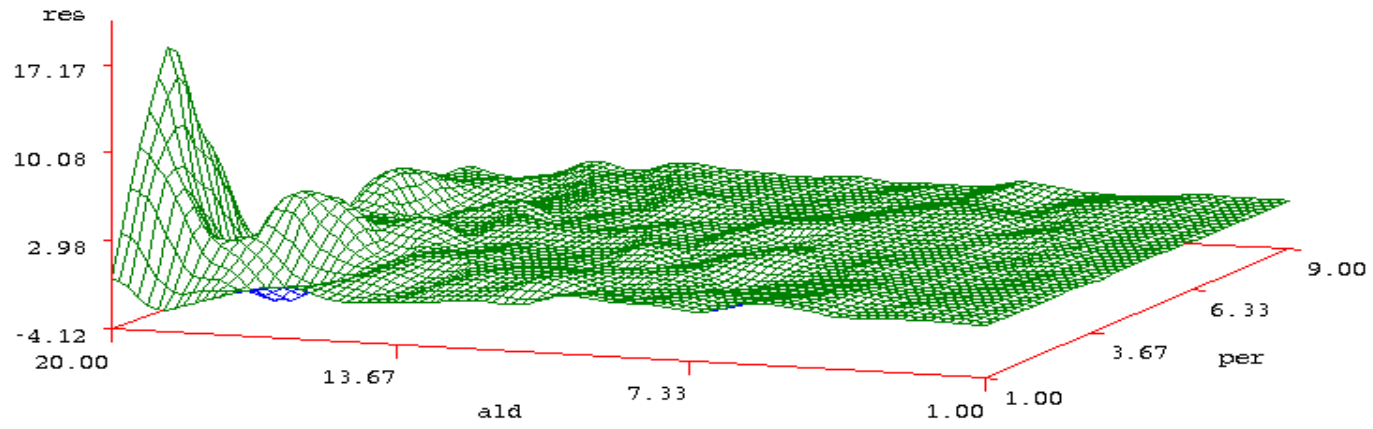
figur 1e

int.5 skattade risker



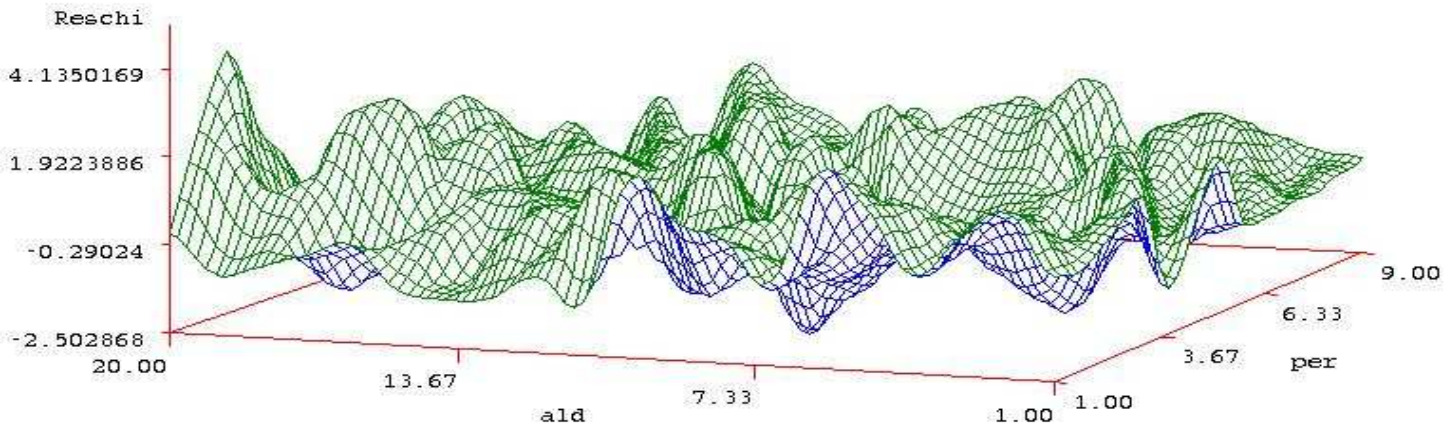
figur 1f

int.5-residualer (obs-pred)



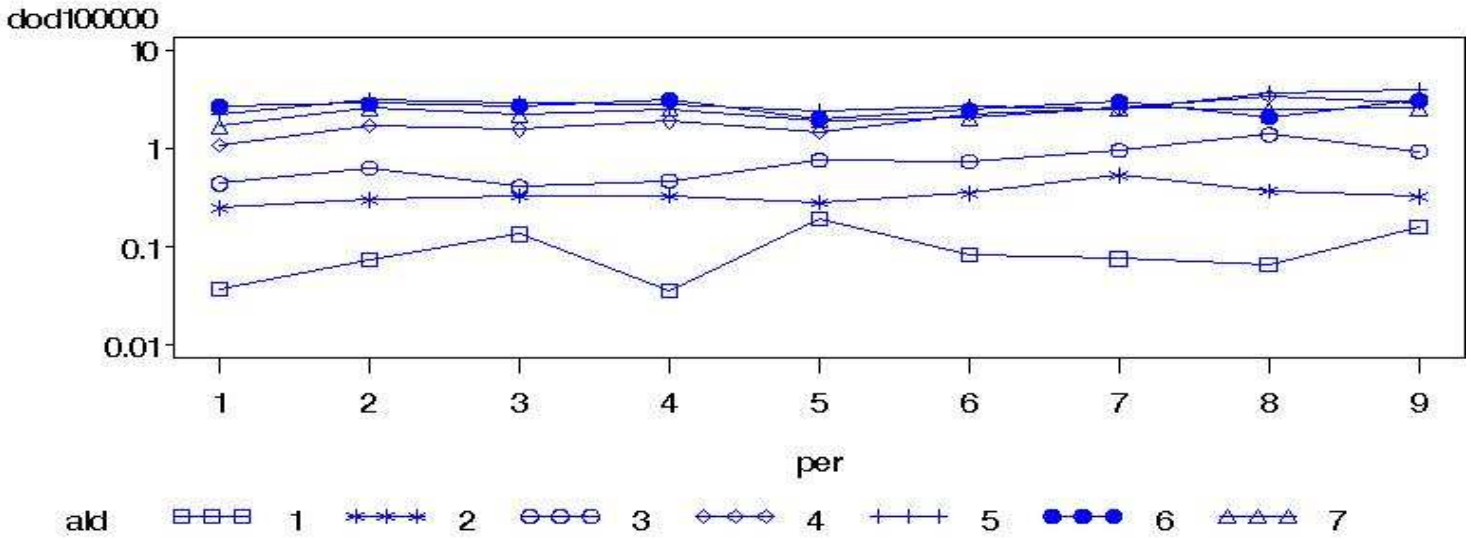
figur 1 g

int=5 ; residualer CHI



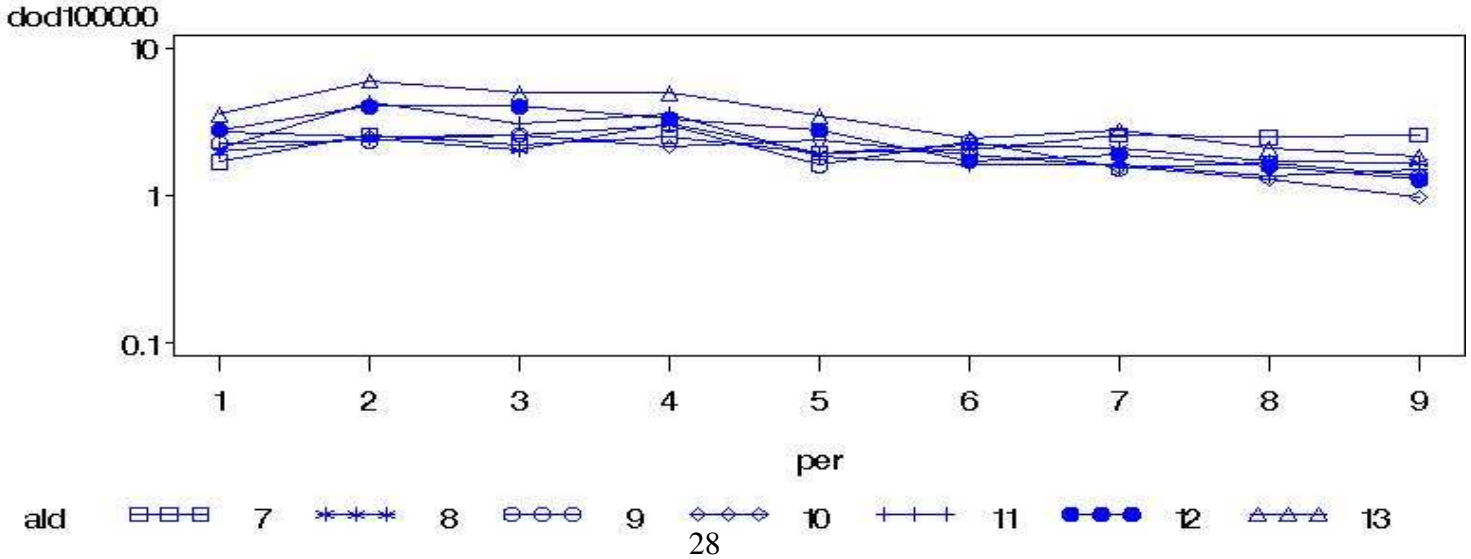
figur 2a

period ald=1–7 Y–axel:log10–skala

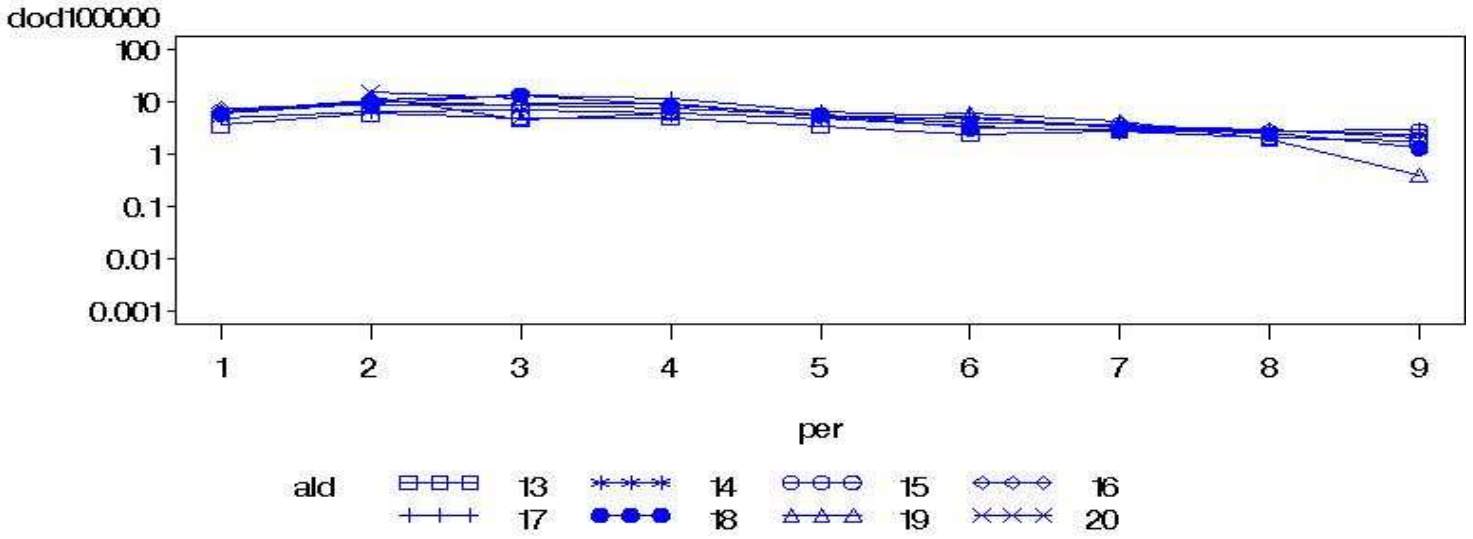


figur 2b

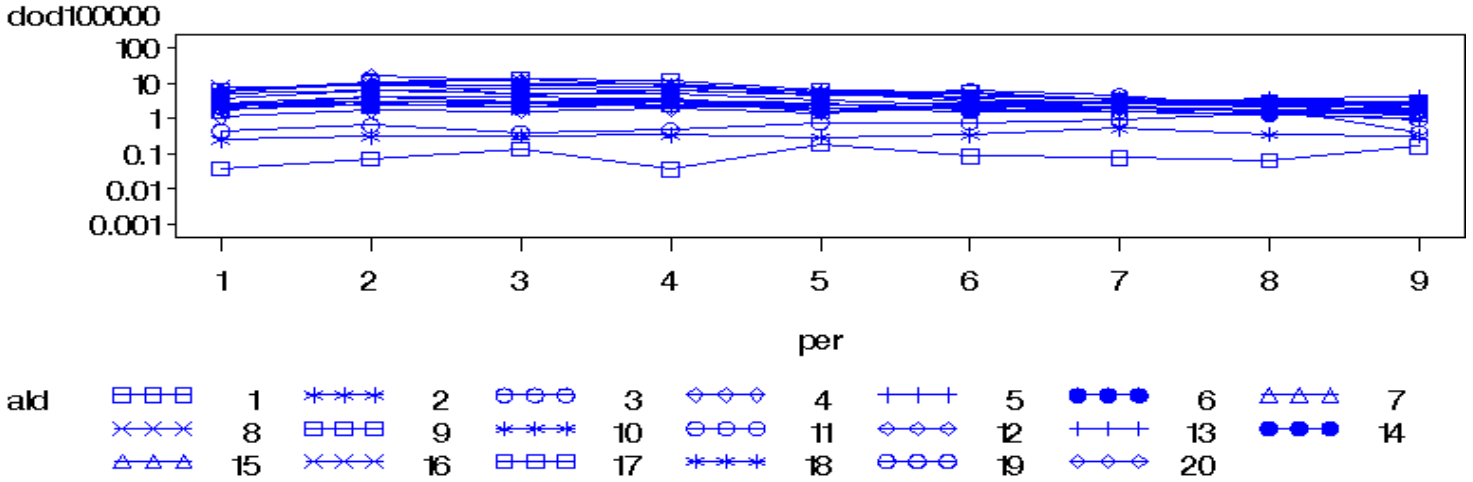
period ald=7–13 Y–axel:log10–skala



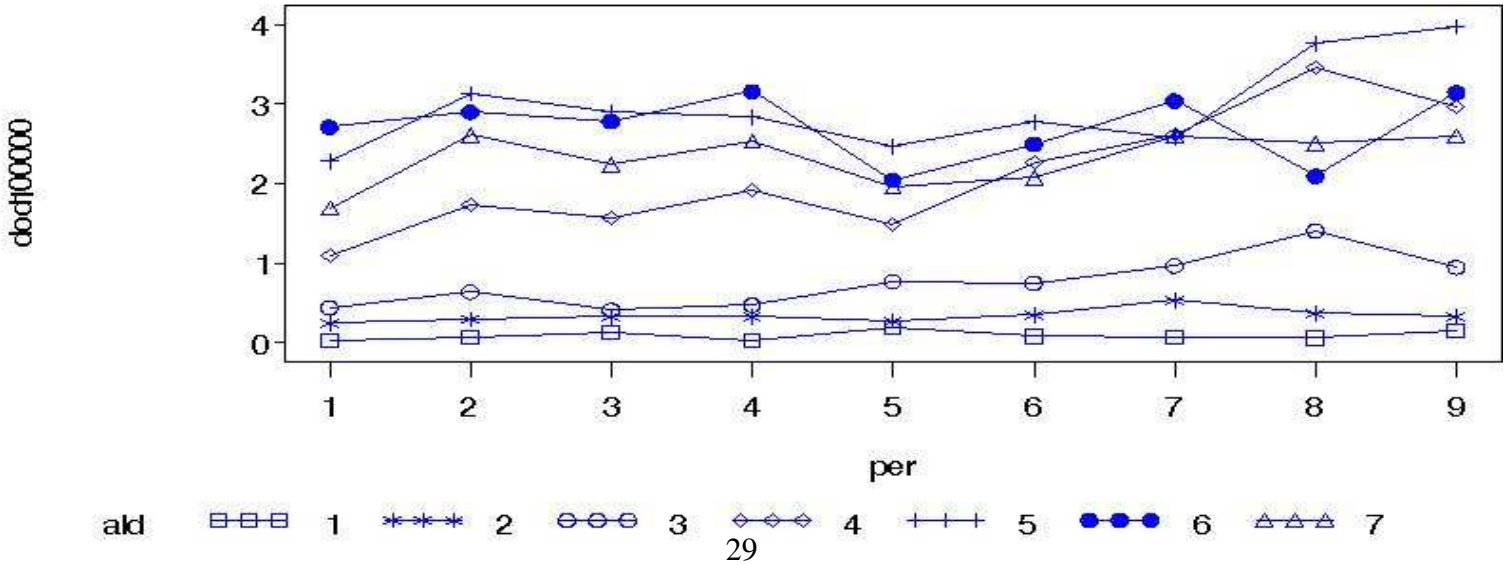
figur 2c
period ald= 13–20 Y–axel:log10–skala



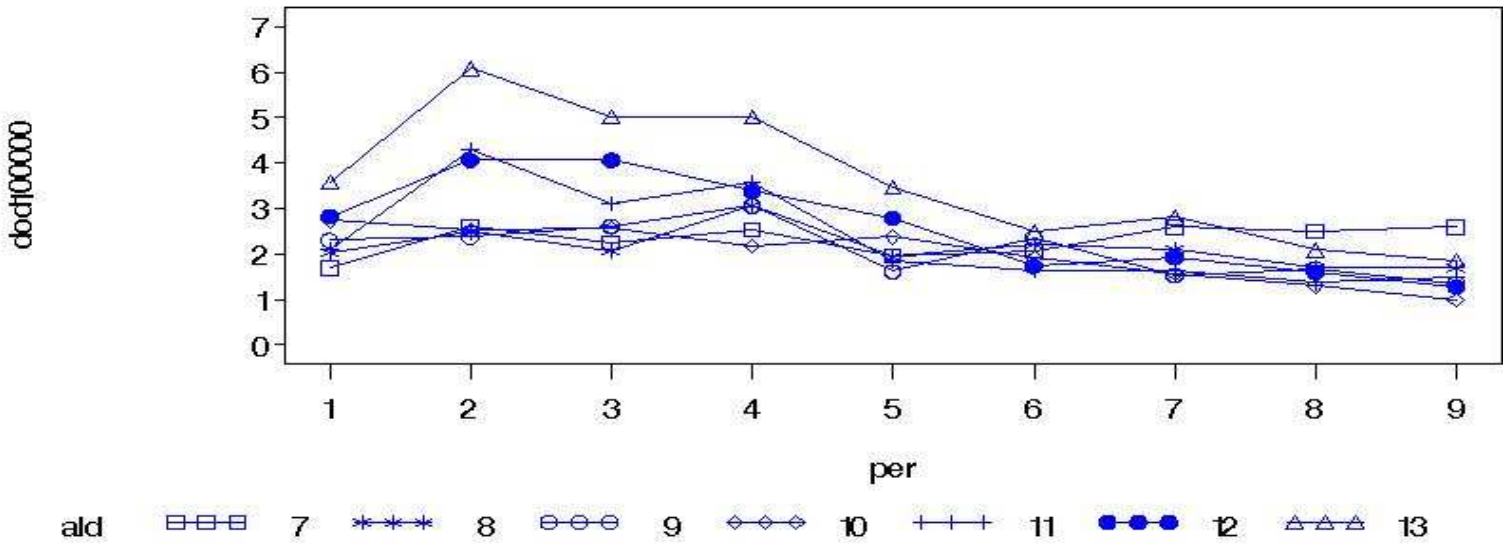
Figur 2 d
Hela datasettet: PERIODER
Y–axel; log–skala



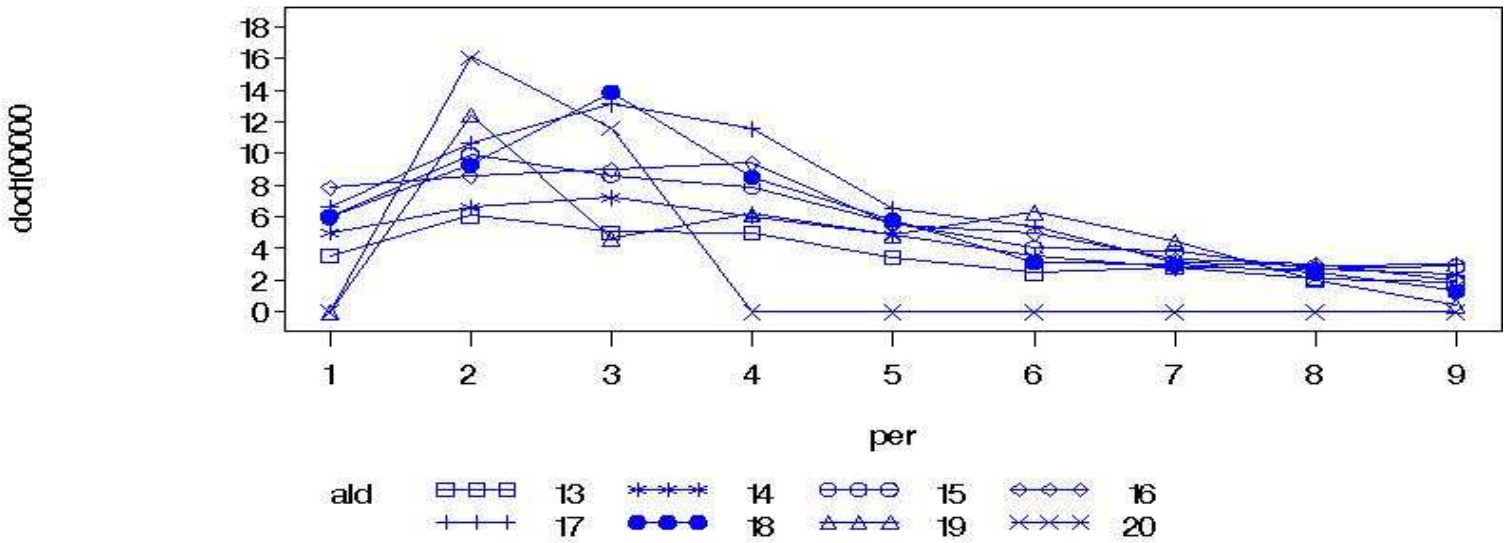
figur 3a
period ald=1–7 Y–axel:normal



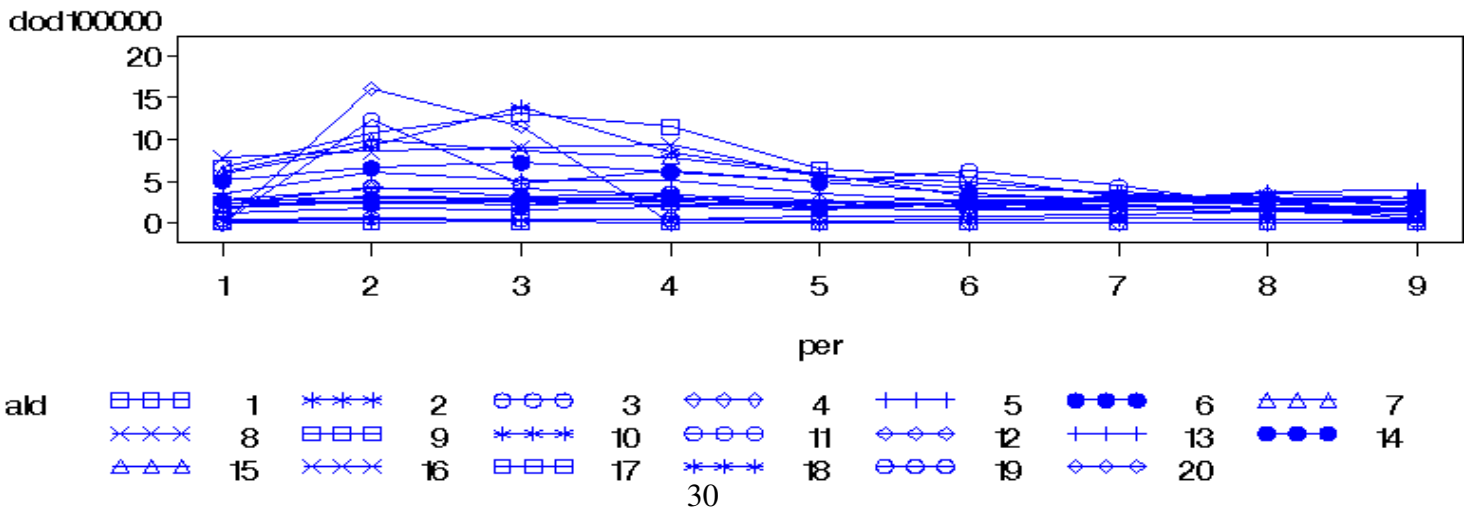
figur 3b
period ald=7–13 Y–axel:normal



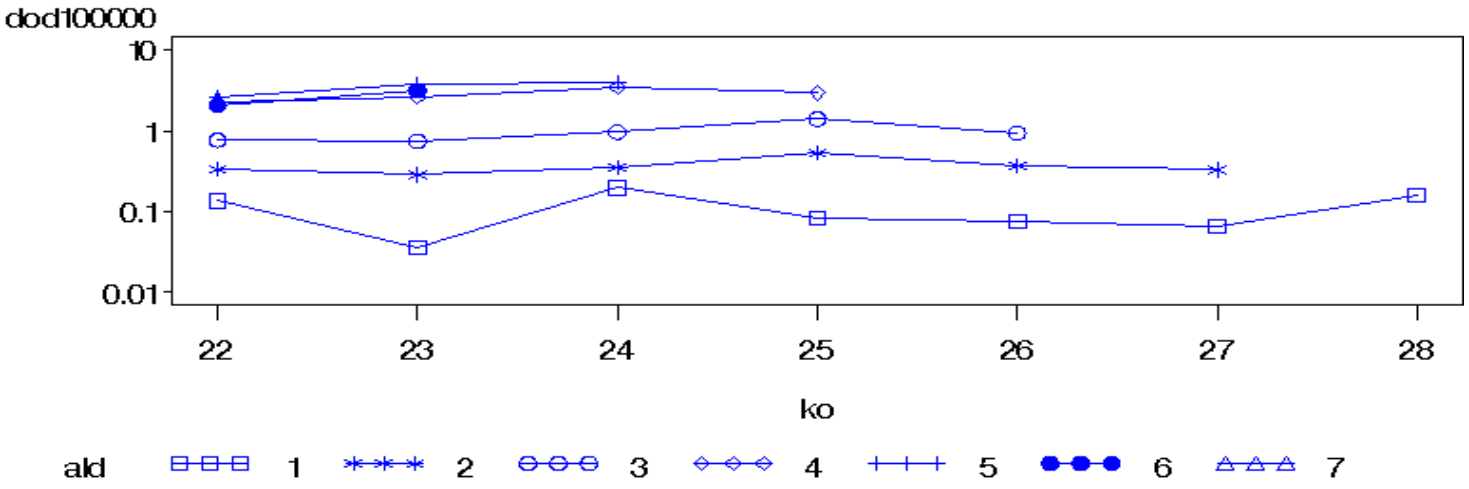
figur 3c
period ald=13–20 Y–axel:normal



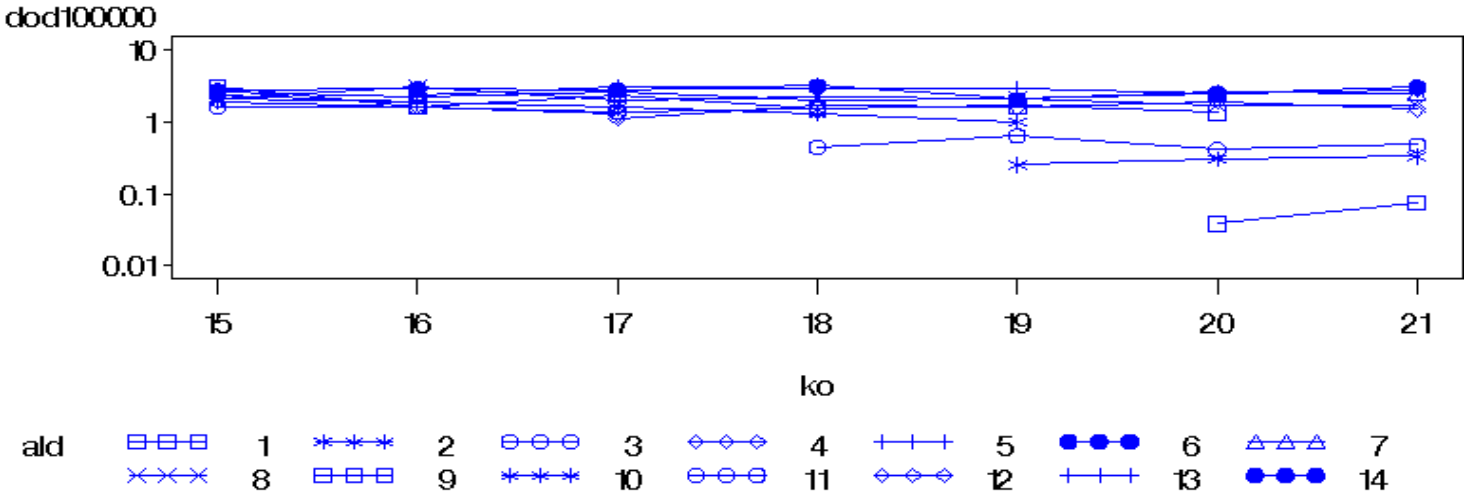
Figur 3 d
Hela datasettet: PERIODER
Y–axel; normal–skala



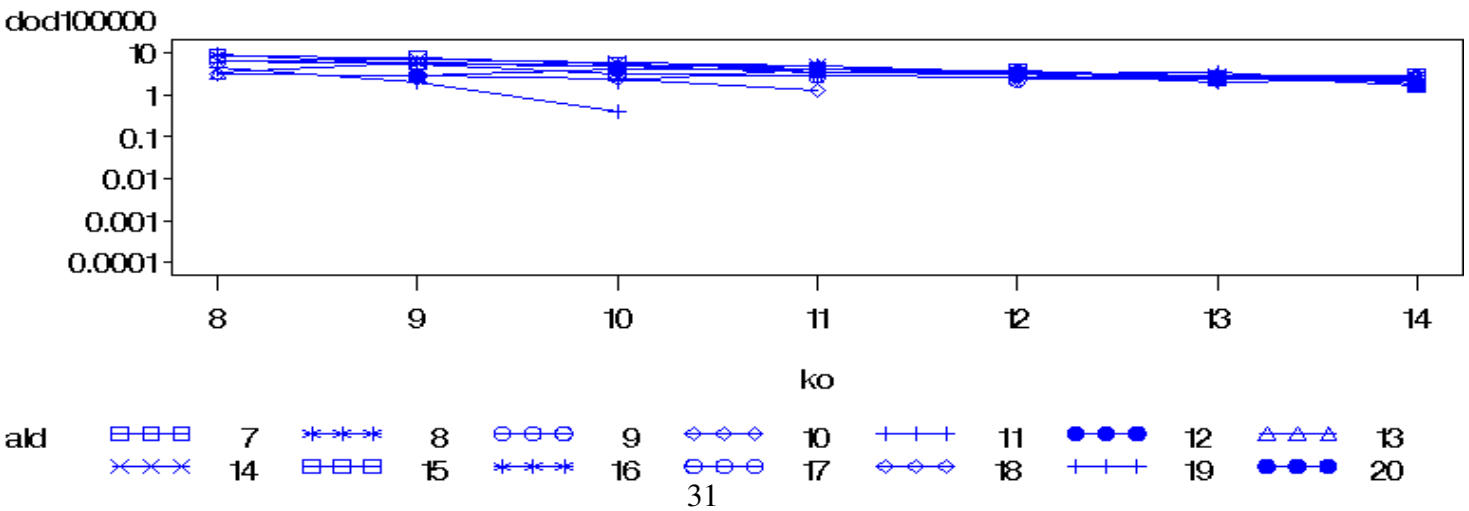
Figur 4 a
22= < kohort
Y–axel : log



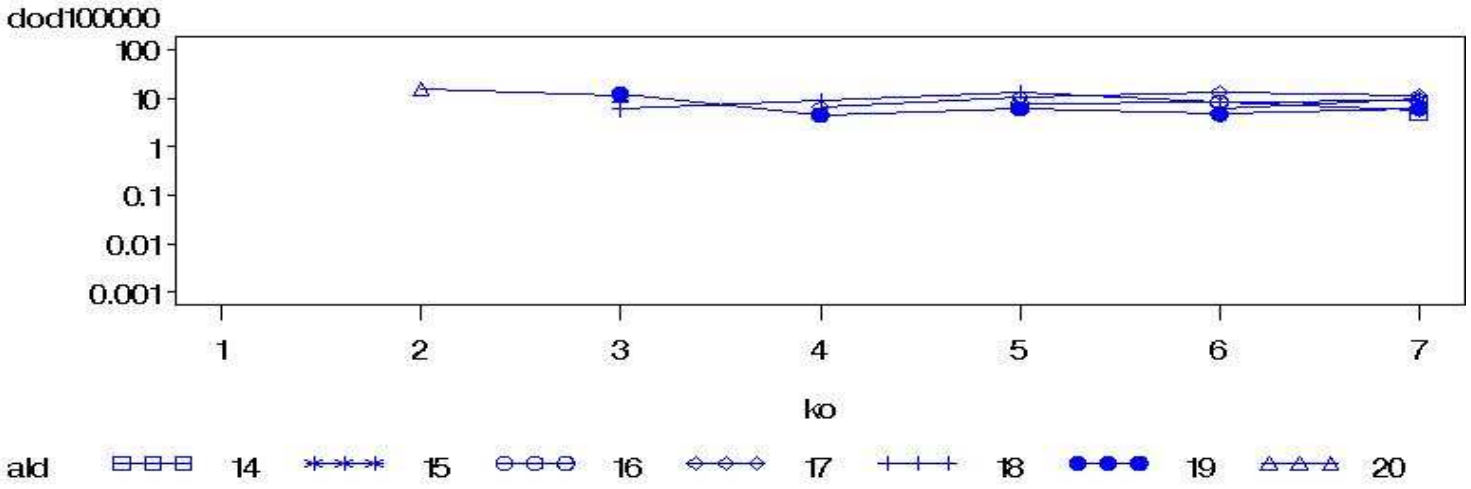
Figur 4 b
15= < kohort= < 21
Y–axel : log



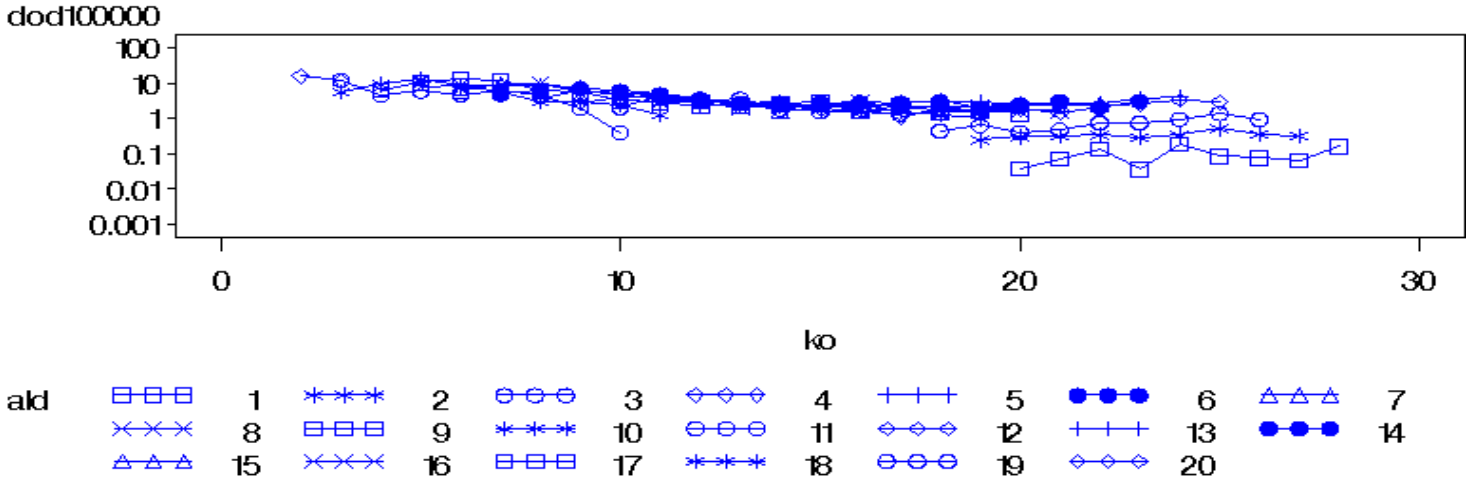
Figur 4 c
8= < kohort= < 14
Y–axel : log



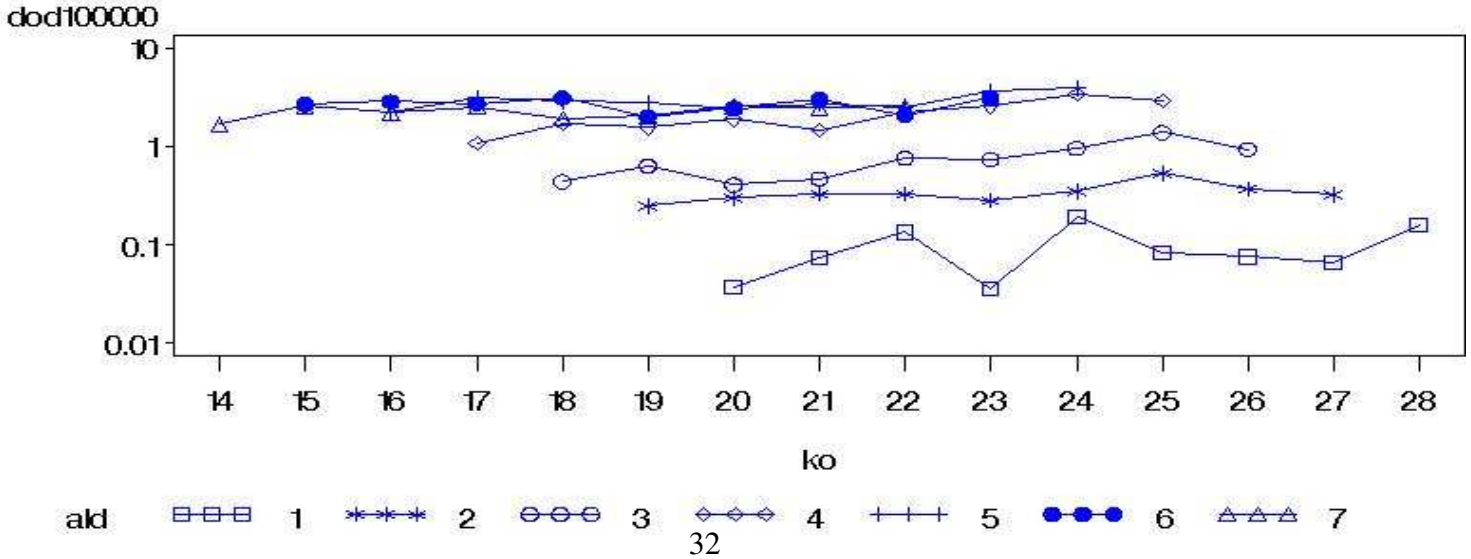
Figur 4 d
kohort= < 7
Y–axel : log



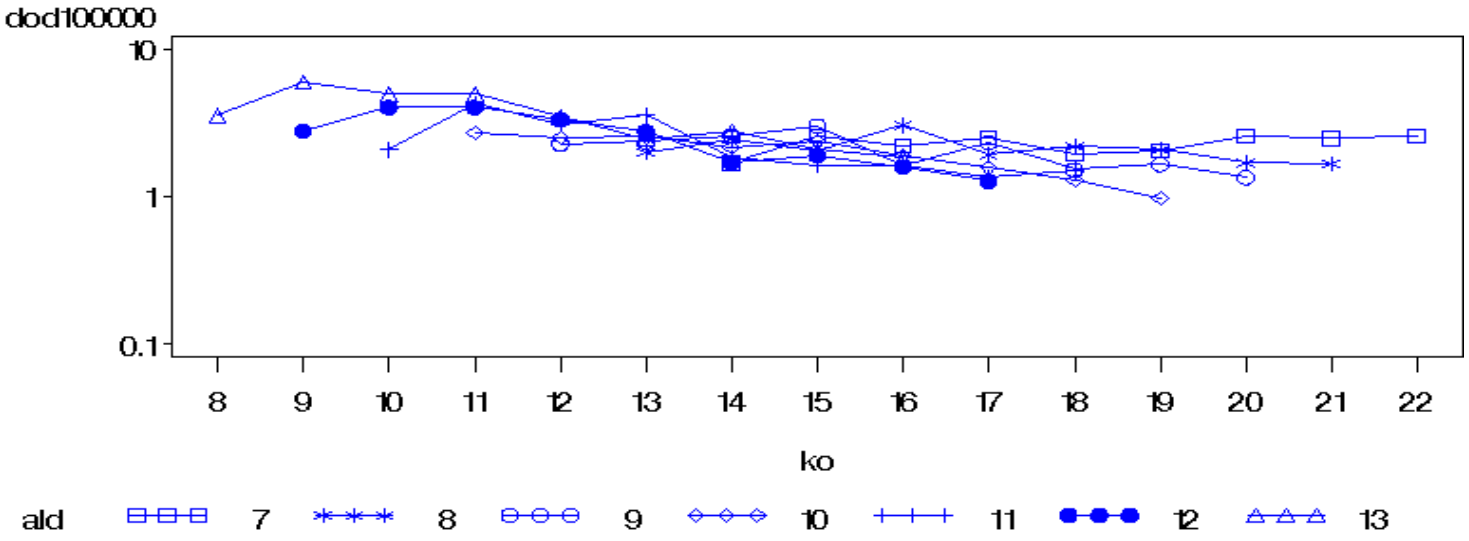
Figur 4 e
Hela datasettet: KOHORTER
Y–axel; log–skala



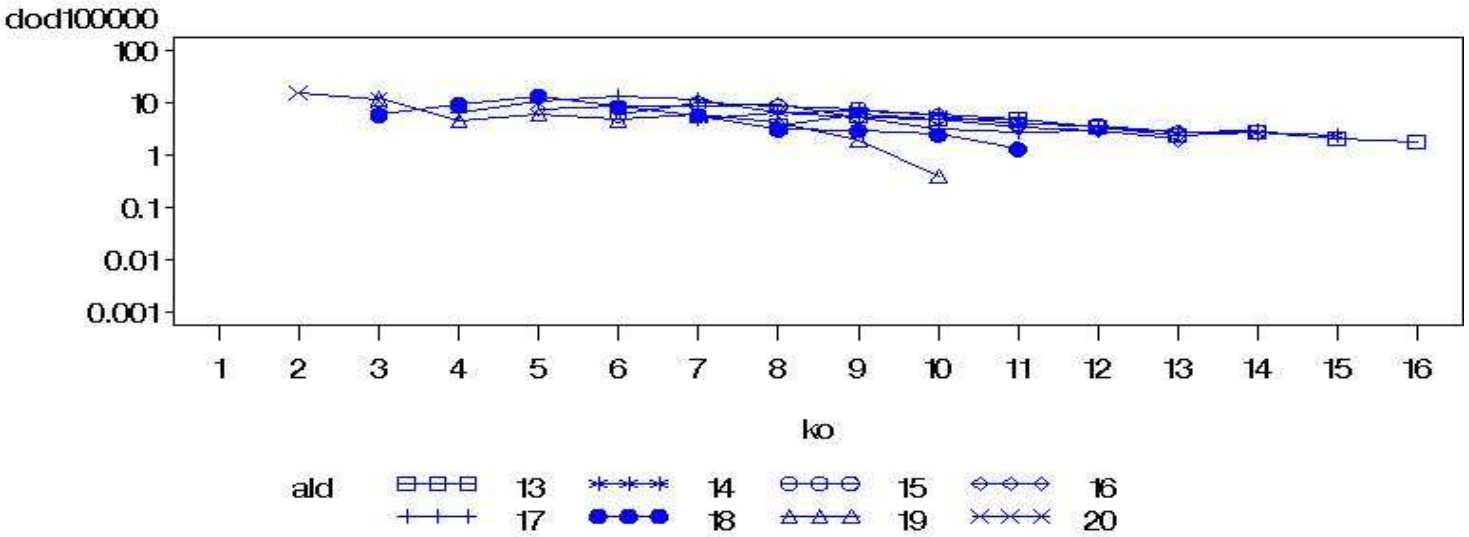
figur 4aa
kohort ald= 1–7 Y–axel:log10–skala



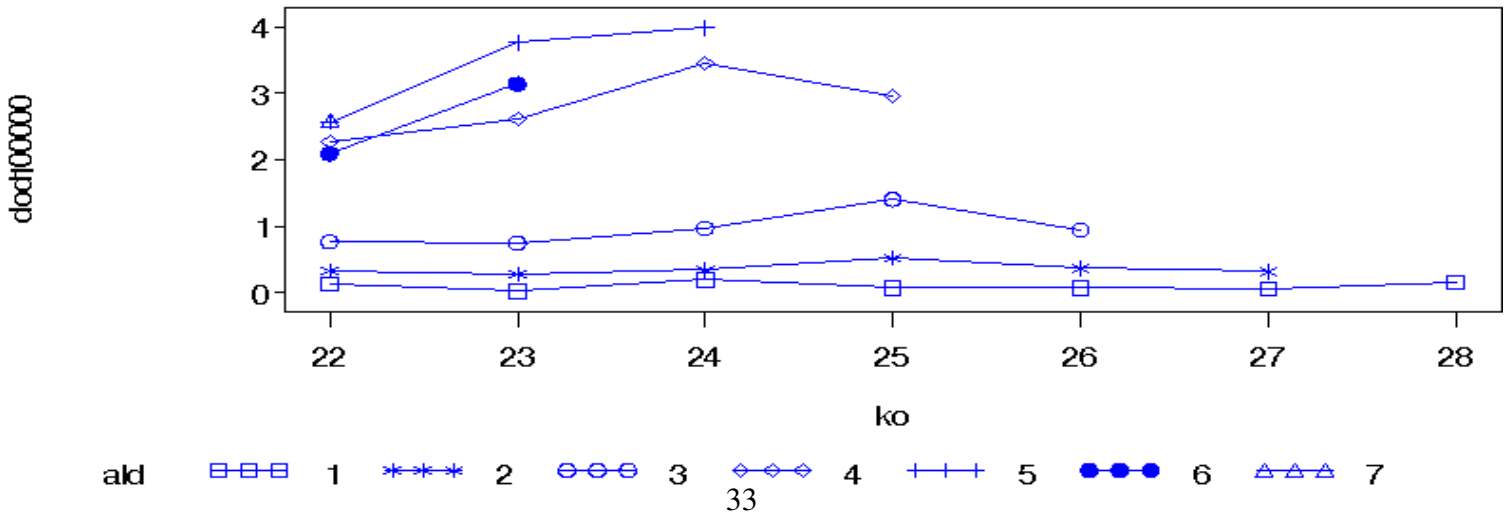
figur 4bb
kohort ald=7–13 Y–axel:log10–skala



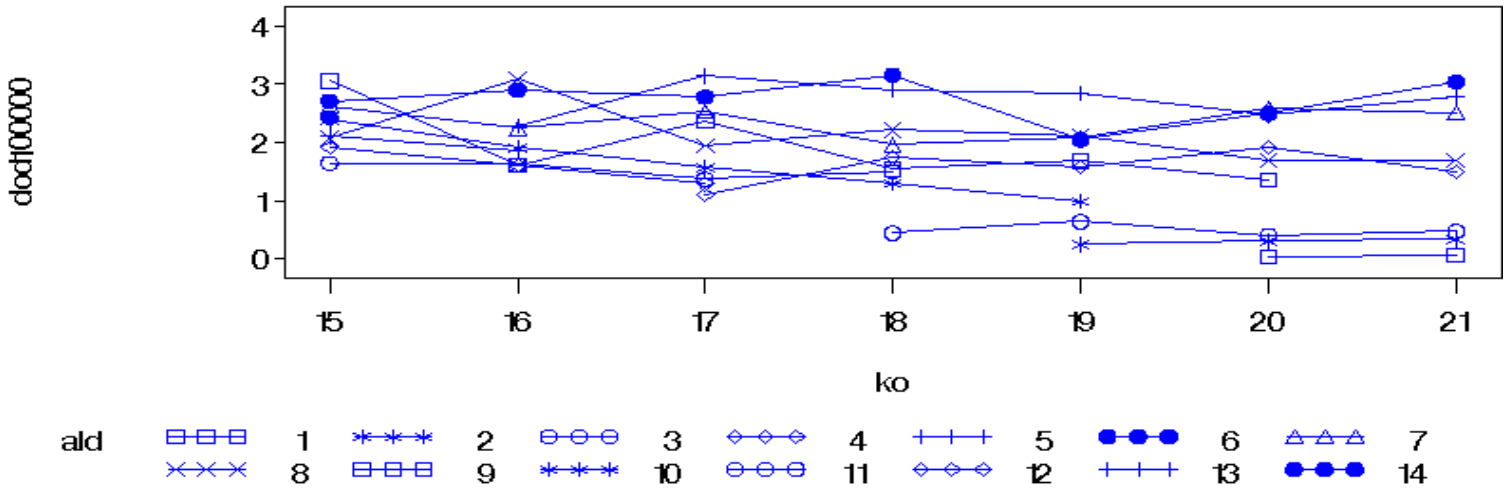
figur 4cc
kohort ald= 13–20 Y–axel:log10–skala



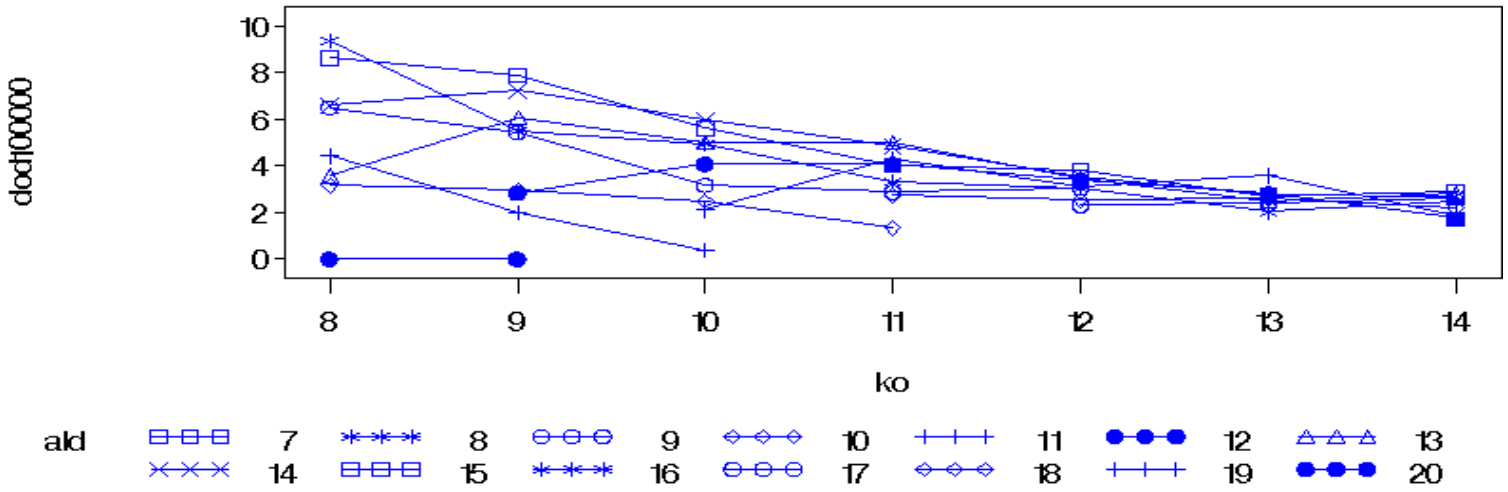
Figur 5 a
22= < kohort
Y–axel : normal



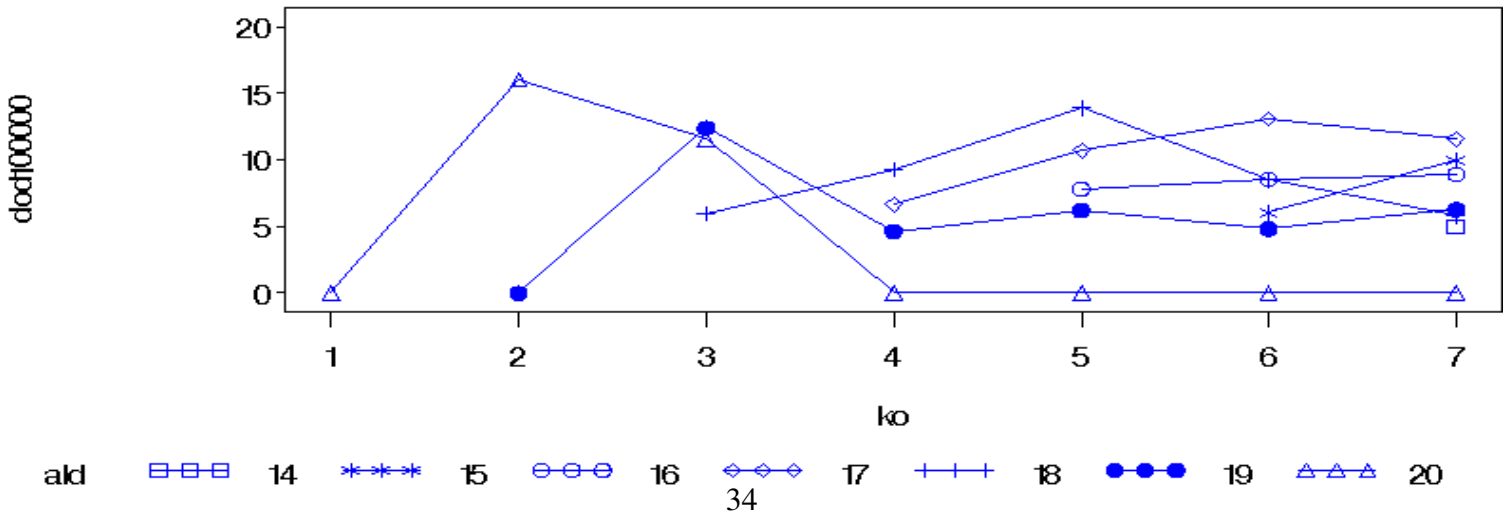
Figur 5 b
15= < kohort= < 21
Y–axel : normal



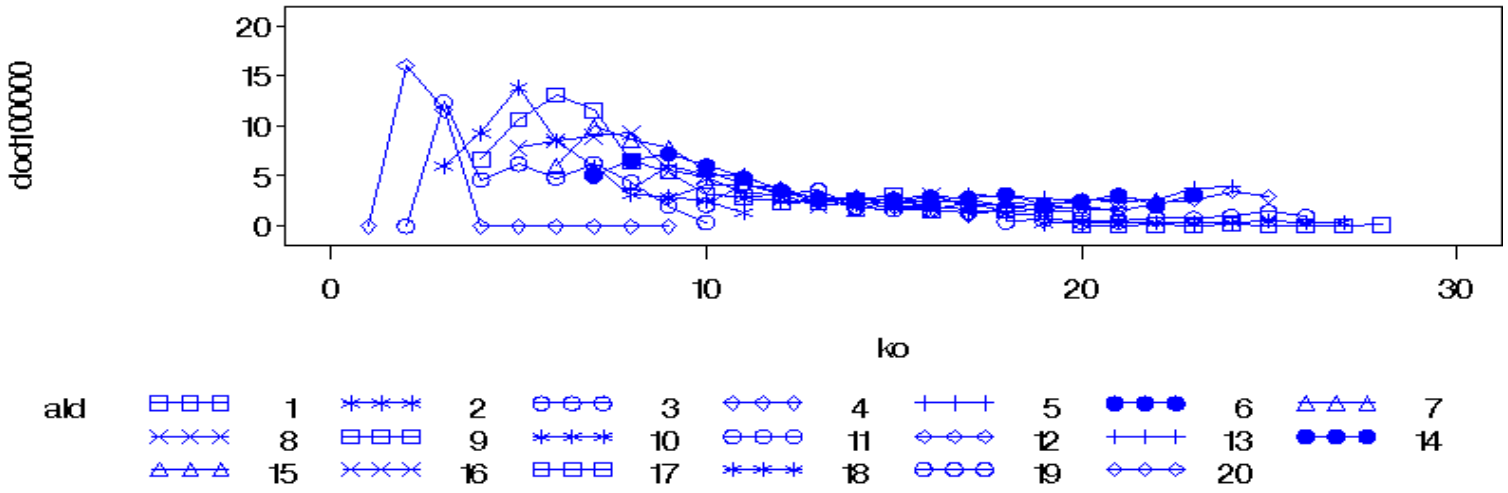
Figur 5 c
8= < kohort= < 14
Y–axel : normal



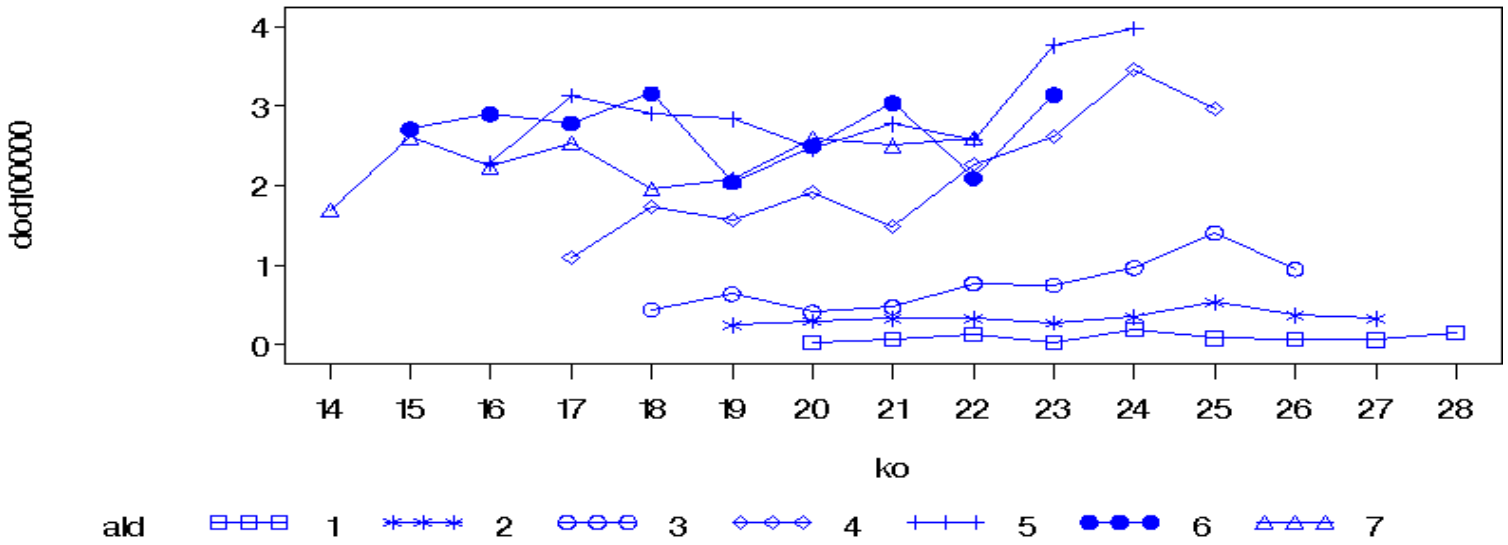
Figur 5 d
kohort= < 7
Y–axel : normal



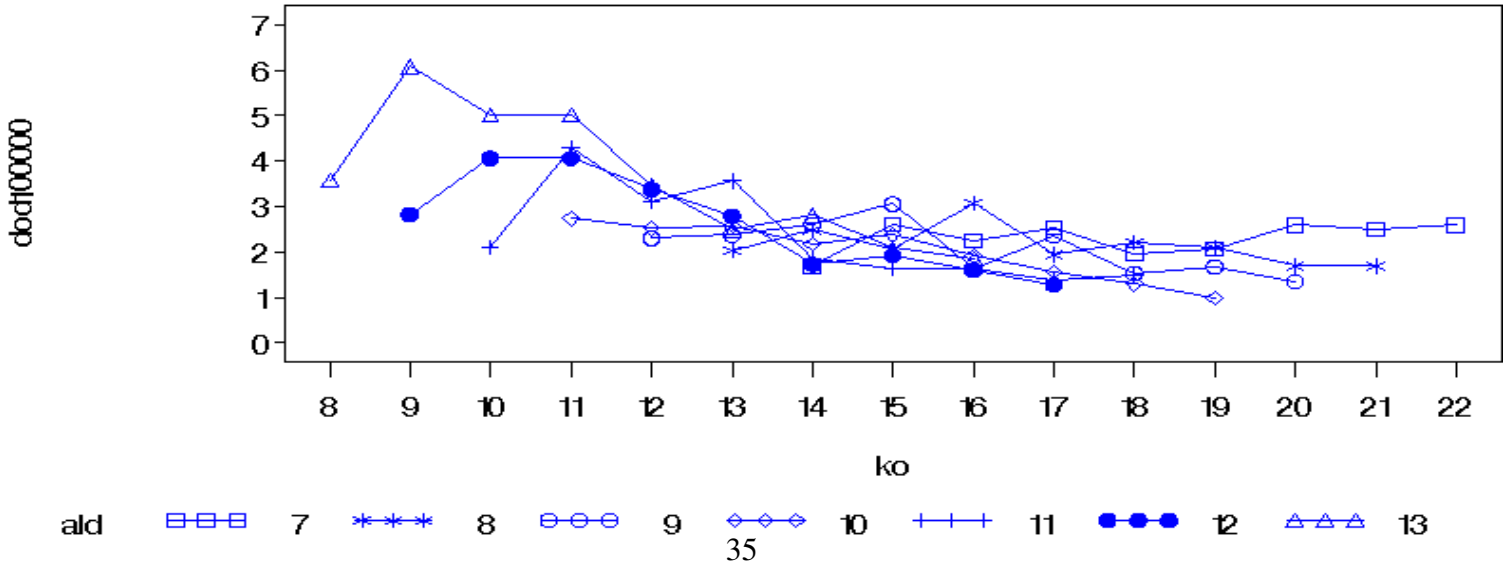
Figur 5 e
Hela datasettet: KOHORTER
Y–axel; normal–skala



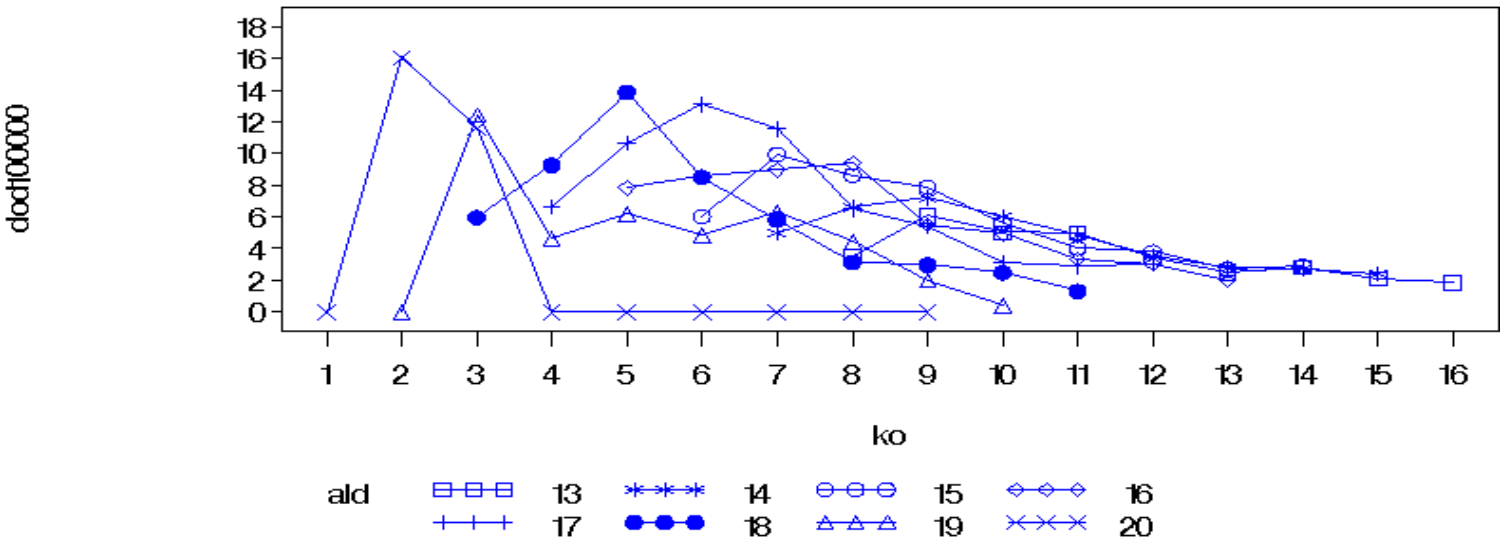
figur 5aa
kohort ald=1–7 Y–axel:normal



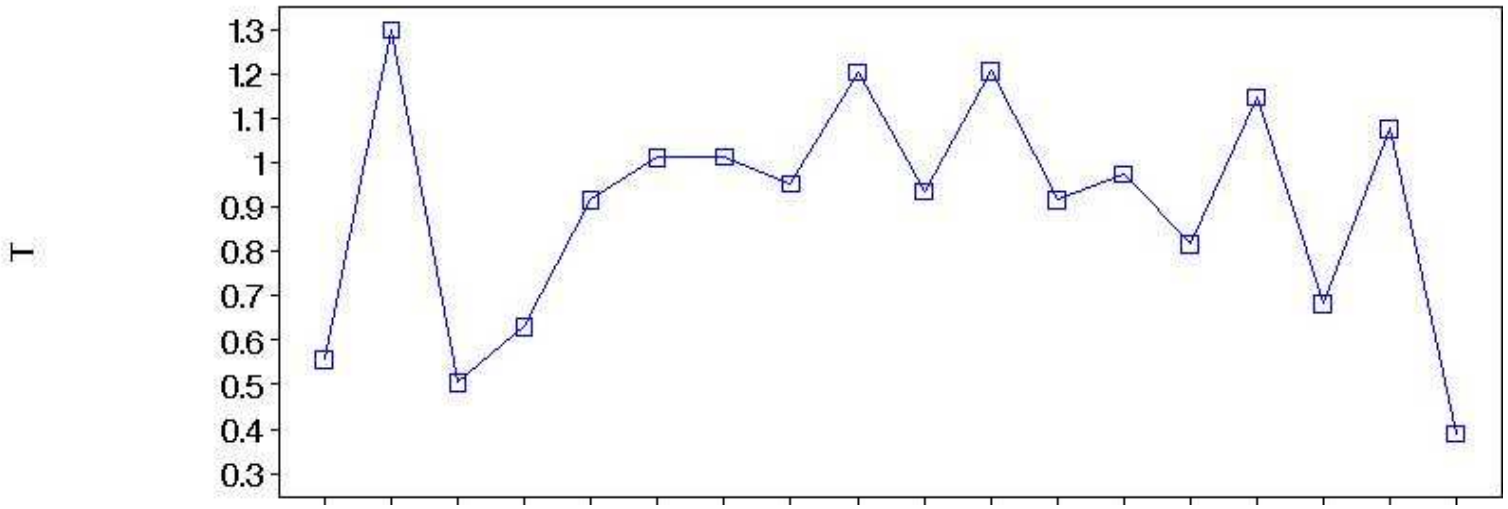
figur 5bb
kohort ald=7–13 Y–axel:normal



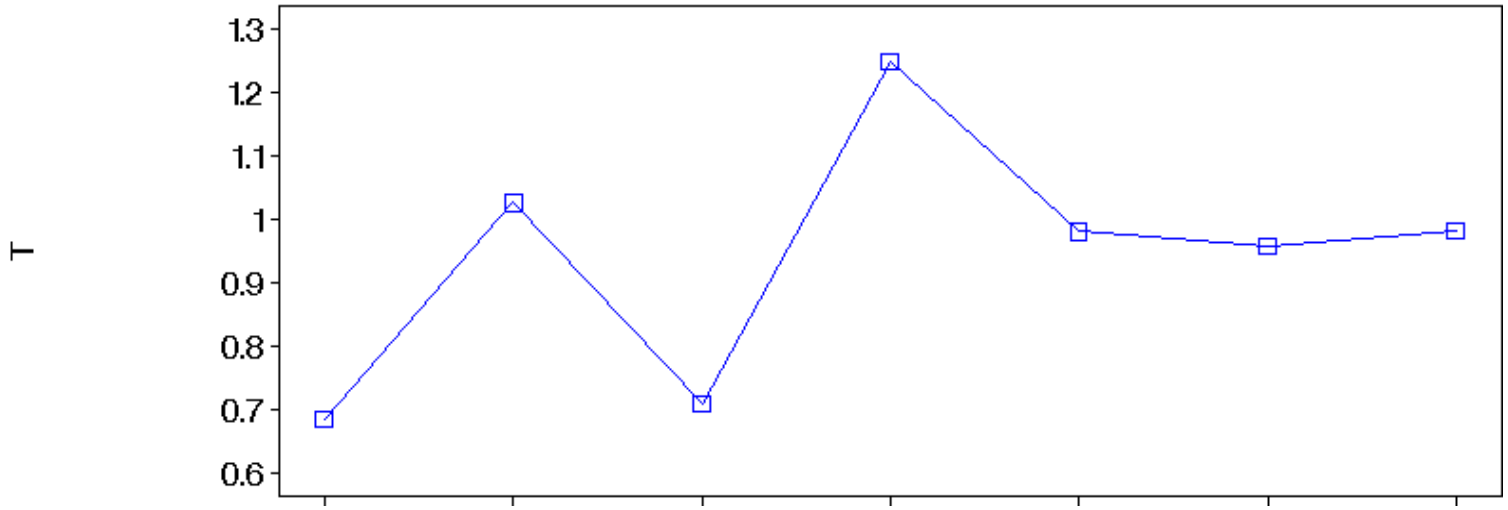
figur 5cc
kohort ald=13–20 Y–axel:normal



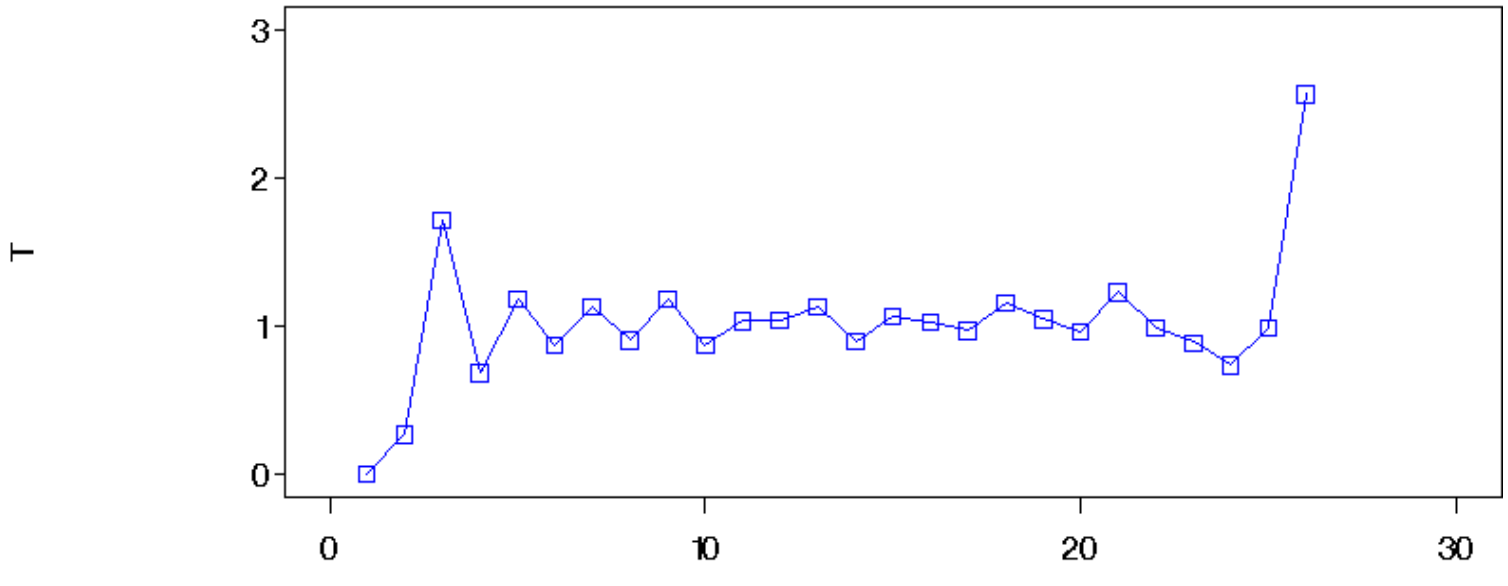
Figur 6a
relativa risker alder int–5



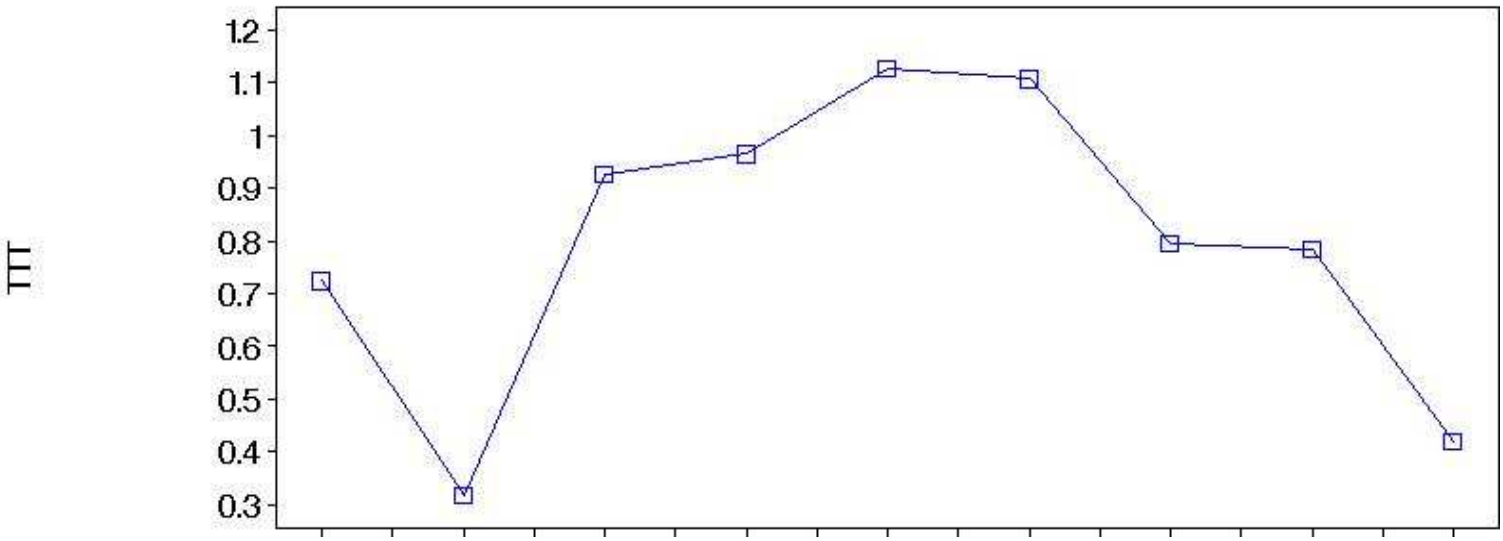
Figur 6b
relativa risker period int–5



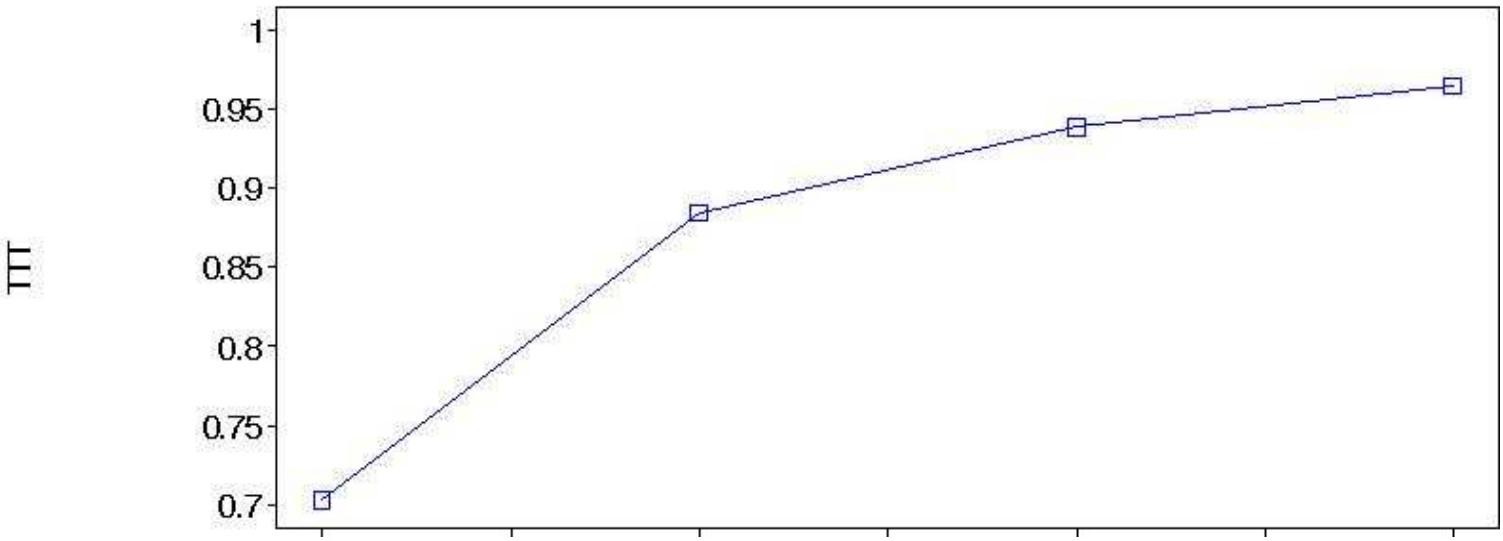
Figur 6c
relativa risker kohorter int–5



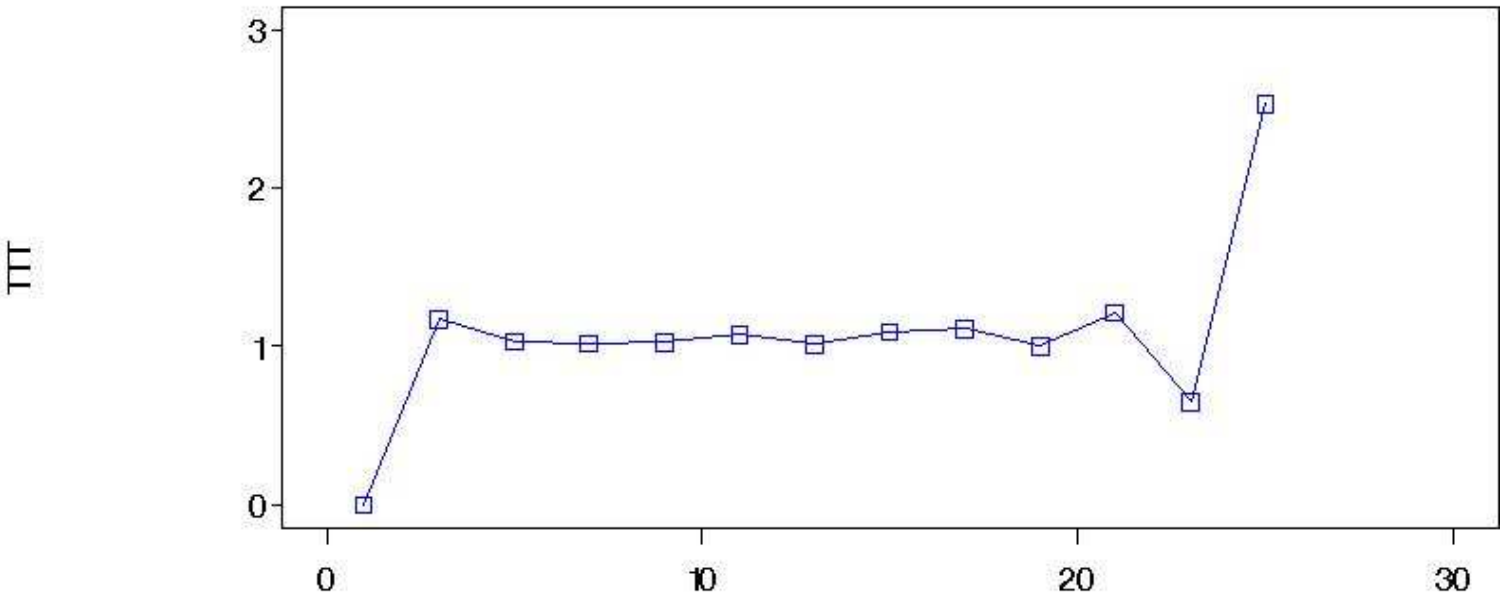
Figur7a



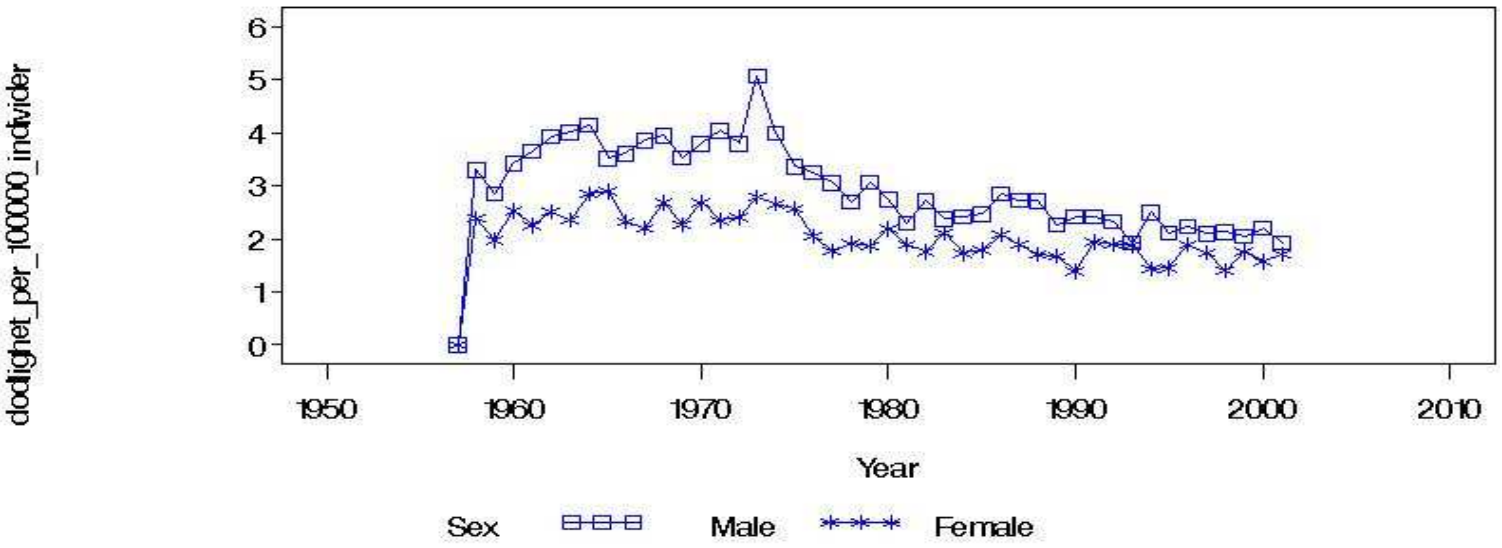
Figur7b



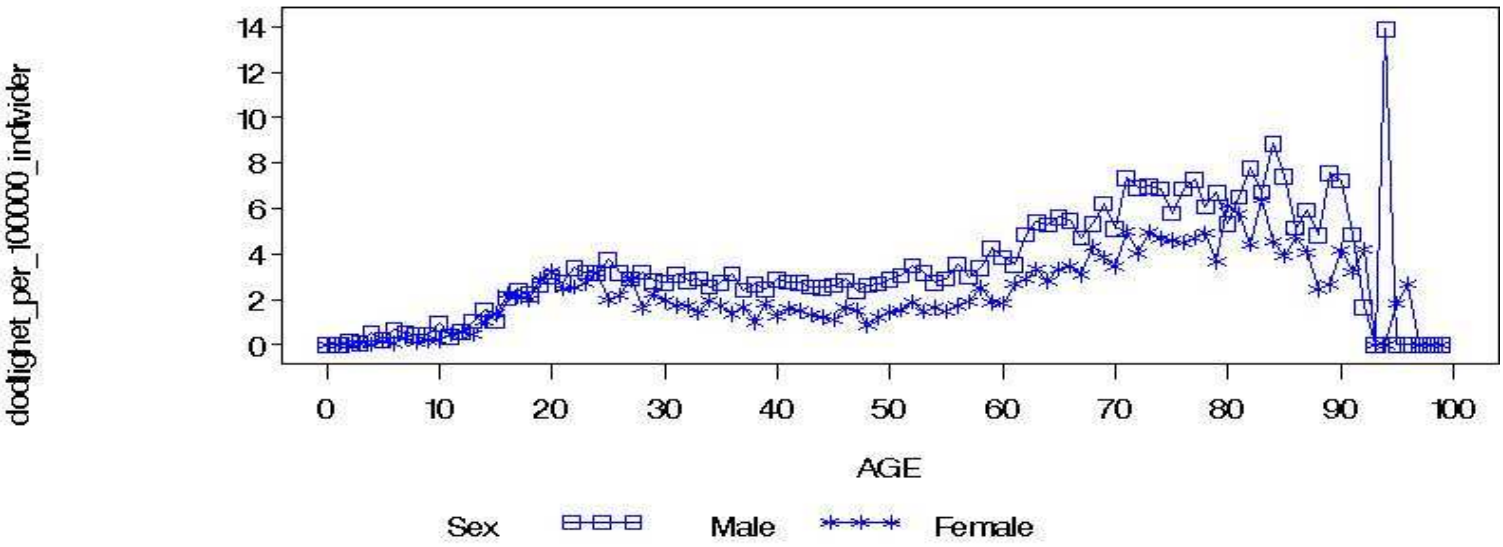
Figur7c



figur 8a
empiriska fall per 100000 individer (oviktat): perioder



figur 8b
empiriska fall per 100000 individer (oviktat): alder



figur 8c
empiriska fall per 100000 individer (oviktat): kohorter