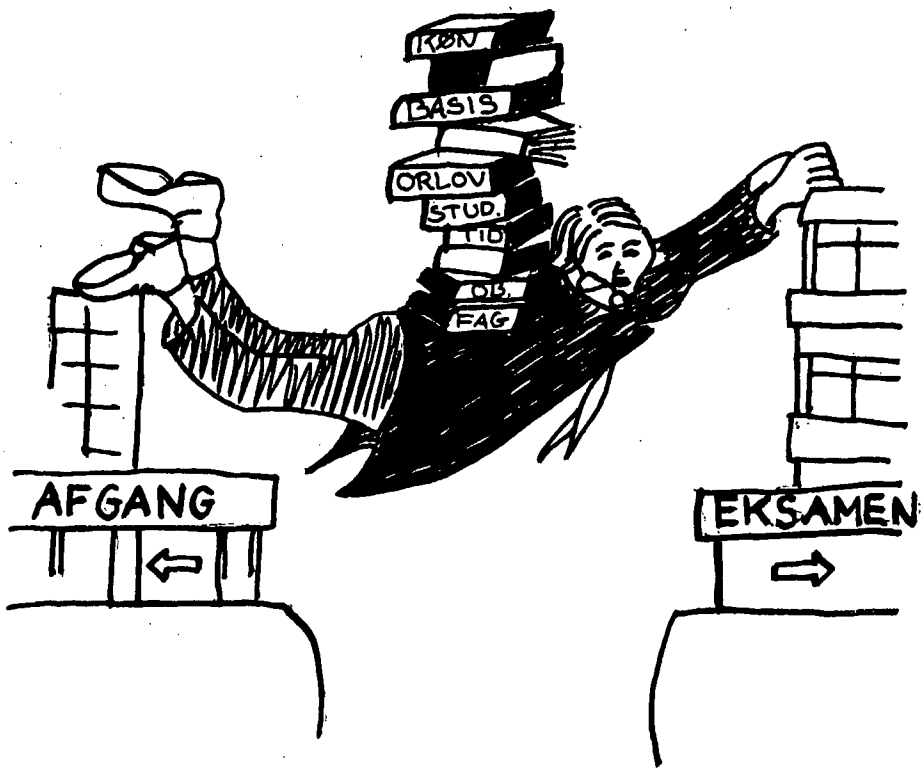
TEKST NR 109 1985

COX I STUDIETIDEN.

- Cox's regressionsmodel anvendt på studenteroplysninger fra RUC.



TEKSTER fra

IMFUFA ROSKILDE UNIVERSITETSCENTER
INSTITUT FOR STUDIET AF MATEMATIK OG FYSIK SAMT DERES FUNKTIONER I UNDERVISNING, FORSKNING OG ANVENDELSER

IMFUFA, Roskilde Universitetscenter, Postbox 260, 4000 Roskilde

COX I STUDIETIDEN - Cox's regressionsmodel anvendt på studenteroplysninger fra RUC.

af Mikael Wennerberg Johansen, Poul Kattler og Torben J Andreasen.

IMFUFA tekst nr. 109/1985 122 sider ISSN 0106 - 6242

Abstract.

Kan man give en fornuftig statistisk beskrivelse af, hvor lang tid der går fra en studerende bliver indskrevet vedRUC til vedkommende består den afsluttende eksamen eller melder sig ud ?

Som første skridt til at belyse dette spørgsmål har vi opbygget en database, hvor enhver der i tidsrummet 1.9.1972 til 1.9.1984 i mindst en måned har været indskrevet som studerende ved RUC er registreret med oplysninger om bl.a. køn, alder, startår, valg af basisuddannelse, valg af overbygningsuddannelse, den totale registrerede studietid, samt hvorvidt vedkommende pr. 1.9.1984 havde taget sin afsluttende eksamen eller havde meldt sig ud uden eksamen eller stadig var studieaktiv.

Dernæst er de indsamlede oplysninger analyseret ved hjælp af Cox's regressionsmodel (Cox 1972), der går ud på at specificere fordelingen af ventetiden indtil afgang med/uden eksamen ved at fortælle, hvordan den tilsvarende afgangs-intensitetsfunktion afhænger af forskellige "baggrundsvariable" som køn, alder, startår, basisuddannelse, overbygningsuddannelse: person nr. j har intensiteten

$$\lambda_j(t) = \lambda_o(t) \exp(\underline{\beta} \cdot \underline{z}_j) \quad ,$$

hvor vektoren $\underline{z}_j$ er en passende kvantificeret udgave af person nr. j 's "baggrundsvariable", $\underline{\beta}$ er en ukendt regressionsparameter, og λ_o er en ukendt funktion.

Projektets titel: COX I STUDIETIDEN.
- Cox's regressionsmodel anvendt
på studenteroplysninger fra RUC.

Deltagere : Mikael Wennerberg Johansen
Poul Kattler
Torben Jens Andreasen

Vejleder : Jørgen Larsen

Matematik overbygningen, forår 1985, modul II.

INDHOLDSFORTEGNELSE.

1. INDLEDNING OG PROBLEMFORMULERING.

Dette projekt tager på alle måder udgangspunkt i RUC. Det er skrevet af RUC'ere, og det drejer sig om RUC'ernes studiemønstre. Projektets bærende idé er om vi med en statistisk metode kan undersøge, hvordan studiemønstret er på RUC og give nogle bud på hvilke studiemæssige forhold, der på afgørende måde præger de studerendes studie på RUC. De nærmere undersøgelser vil især dreje sig om hvilke studier og studieforhold, der giver de studerende den bedste mulighed for at få eksamen og på hvilken tid.

Til analysen anvender vi Cox's model for overlevelsesdata. Modellen tager sit udgangspunkt i oplysninger (data), som man får fra hver enkelt person, der indgår i undersøgelsen. De anvendelsesområder, hvor man hidtil har brugt modellen, har været inden for lægevidenskaben. Her har man især anvendt modellen ved kræftpatienter og deres levetider indtil døden, men også inden for biologien er modellen hyppigt anvendt, eksempelvis kan man studere forsøgsdyrs levetider ved at eksponere disse med mistænkte toxikologiske stoffer. Til trods for, at man i overlevelsesanalyse, som navnet antyder, især beskæftiger sig med undersøgelser, hvor man betragter levetider indtil døden indtræffer, skal det pointeres, at modellen har et langt bredere anvendelsesområde. Godt nok har vi ikke stødt på mange anvendelser af modellen, hvor slutresultatet ikke har været døden, men så kan projektet være en yderligere dokumentation for modellens generelle anvendelighed.

Det resultat, som Cox-modellen giver,er en beskrivelse af overlevelsen som funktion af tiden, indtil den undersøgte hændelse indtræffer; og i vores tilfælde bliver det eksamen eller udmeldelse (afgang uden eksamen). For alle de studerende, der har været på RUC siden 1972, har vi således indsamlet deres personlige

data omkring studieforløbet. Disse personlige data er bl.a. studiestart, alder, køn, overbygningsfag (kaldet OB-fag), basisfag osv. Meget ofte vil disse personlige data blive refereret som en persons variable.

Cox-modellen bruger hver enkelt persons variable til at lave en model over en hel population uanset hvor stor en gruppe personer, der har ligget til grund for modellens udformning. De enkelte personers persondata bliver bearbejdet i en regressionsmodel, der evner at bearbejde personoplysningerne sådan, at de variables betydning for "dødeligheden i patientmaterialet" udtrykkes v.hj.a. nogle parameterværdier.

Modellen opererer inden for den generelle terminologi for overlevelsesanalyser; sammenligner dødeligheder, beregner overlevelsessandsynligheder m.v. Dens vigtigste fortrin er imidlertid dens evne til at finde tendenser i en ikke-homogen befolknings overlevelsesmønstre.Hidtil oversete detailoplysninger fra patientgruppen bliver påpeget af modellens parameter-udregninger og symptomers indbyrdes betydning vægtes hermed. Den færdige model er formuleret i generelle vendinger, og selv om den er beregnet udfra et bestemt patientmateriale, udtaler den sig også om andre patienters overlevelseschancer med samme sygdom. Hér rummer Cox-modellen nogle oplagte prognose-egenskaber.

Arbejdet med projektet har taget et år, og dette år har været inddelt i fem perioder, hvori vi har arbejdet med større eller mindre delområder af projektet.

I første periode har vi sat os ind i den grundlæggende teori omkring overlevelsesdata og opstillingen af Cox-modellen. Vores referencer har især været Kalbfleisch & Prentice 80 og Andersen og Væth 84.

I den anden periode har vi arbejdet med planlægningen og opbygningen af databasen, samt ansøgt om tilladelse til at bruge RUC's matrikel oplysninger. Imens lavede vores vejleder et COMPAS program til registrering

af de mange data.

I tredie periode arbejdede vi igen med den grundlæggende teori, dog som vejledere på et natbasis kursus i statistik, hvor vi i en måned gennemgik overlevelsesdata som en case,så de studerende kunne få et indtryk af, hvad man kan bruge statistik til.

I fjerde periode satte vi os lidt ind i programmeringssproget COMPAS PASCAL, så vi ved egen hjælp kunne lave små simple statistikker fra vores database og samtidig redigere i den, så den passede til opbygningen af den endelige database. Vi arbejdede også med nogle forskellige emner inden for den mere dunkle teori, såsom censureringsproblemer i forbindelse med konkurrerende afgangsårsager, i vores tilfælde eksamen og afgang uden eksamen.

Sidste periode har været opstilling af den endelige model, som blev foretaget med det statistiske program BMDP på RECKU.

Dét, vi vil, er:

"Vi vil bearbejde Cox-modellen, så vi vil være i stand til at benytte den i en konkret sammenhæng. Vi vil indsamle nye og relevante data fra andre steder end de traditionelle lægevidenskabelige områder. Med udgangspunkt i det indsamlede materiale (RUC-studenternes studieforløb) vil vi afprøve og ombearbejde Cox-modellen udfra et ønske om virkelig at nyttiggøre en del af den matematisk/statistiske teori.

Kan vi med Cox-modellen anvendt på dette felt bidrage med ny viden og stimulere en debat om resultatet?"

2. INTRODUKTION AF OVERLEVELSESDATA.

I dette kapitel vil vi starte med nogle historiske betragtninger omkring overlevelsesdata, og derefter forklare de mest basale begreber, der ligger til grund for forståelsen af denne del af statistikken.

a. Lidt historie om overlevelsesdata.

Vi skal tilbage til omkring 1500 tallet for at finde de første forsøg på at indsamle data og bearbejde disse. Det har først og fremmest været myndighederne, som kunne have en interesse i at have oplysninger om befolkningens tilstand, ligesom det har været nyttigt at have kendskab til ens militære styrke. Man kunne også sagtens forestille sig, at det kunne være bekvemt for skatteindkrævere at have kendskab til befolkningens sammensætning, så kronen og kirken kunne få mest ud af befolkningen både som skatteydere og fæstebønder.

I 1662 udgav en englænder ved navn Graunt en bog, som vel nok er den første egentlige befolkningsstatistik. Selv om han havde gjort sig nogle systematiske og fornuftige overvejelser omkring mangler i fødsels- og dødslisterne, var det først og fremmest ved hjælp af forskellige sjus, at han bestemte et tal for Londons befolkning og lavede en egentlig dødstavle, som vi ser her:

Figur 2.1. Graunts dødstavle.

Alder	Antal døde i procent	Alder	Antal overlevede
0- 6	36	6	64
6-16	24	16	40
16-26	15	26	25
26-36	9	36	16
36-46	6	46	10
46-56	4	56	6
56-66	3	66	3
66-76	2	76	1
76-86	1	86	0

Westergaard, 1931, s. 22.

Hvordan Graunt egentlig lavede sin statistik og hvordan man skal tolke resultaterne må nok stå lidt hen i det uvisse, men forsøget er interessant,for allerede her har vi ideen til en type statistik, hvor man i generelle vendinger kan sige noget om befolkningens liv og død, når man lader enkeltindividet ude af billedet og betragter en større befolkning som en levende mekanisme. Graunt gjorde sig nogle på forhånd fastlagte antagelser; f.eks. var et menneskes middellevetid konstant, og han kendte ikke noget til befolkningens sammensætning på køn og alder, før han lavede sine tabeller.

Andre folk (ikke matematikere) regnede også på befolkningstal. Udover at tælle huse og husstande og gange det med en formodet husstandsstørrelse, kunne man beregne folketallet udfra den nogenlunde stabile rate fødselstallet dengang udgjorde. I en fransk undersøgelse fra 1766 (se Westergaard, 1931, s. 79-80) beregner man udfra en mindre undersøgelse, at der årligt fødes et barn for hver 25 individer, og da barnefødslerne opgjordedes i kirkebøgerne, kunne man nu ganske simpelt multiplicere antallet af aktuelle fødsler i et bestemt år med 25 og så havde man befolkningstallet.

Den første, som helt præcist beskæftigede sig med overlevelsesstudie, var Daniel Bernouilli, der i 1766 udgav en bog om den frygtede sygdom koppers virkninger i befolkningen i tilknytning til en nyopfunden vaccine. Det interessante var selvfølgelig dengang at undersøge, om hvorvidt der var nogen fordele ved at lade sig vaccinere (hvilket mange tvivlede på). På trods af et dårligt talmateriale udmærkede Bernouilli's arbejde sig ved, at han indførte brugen af kontinuerte funktioner i dødelighedsterminologien, og når overlevelseskurven var kontinuert, kunne man pludselig regne på chancen for at gå hen og dø af kopper; nemlig v.hj.a. kurvens hældning og øvrige form. Bernouilli inddrog begreber, som vi nu kender fra funktionsanalyse i matematik og løsning af differentialligninger i al snakken om dødelighed og sammenligning af dødelighed.

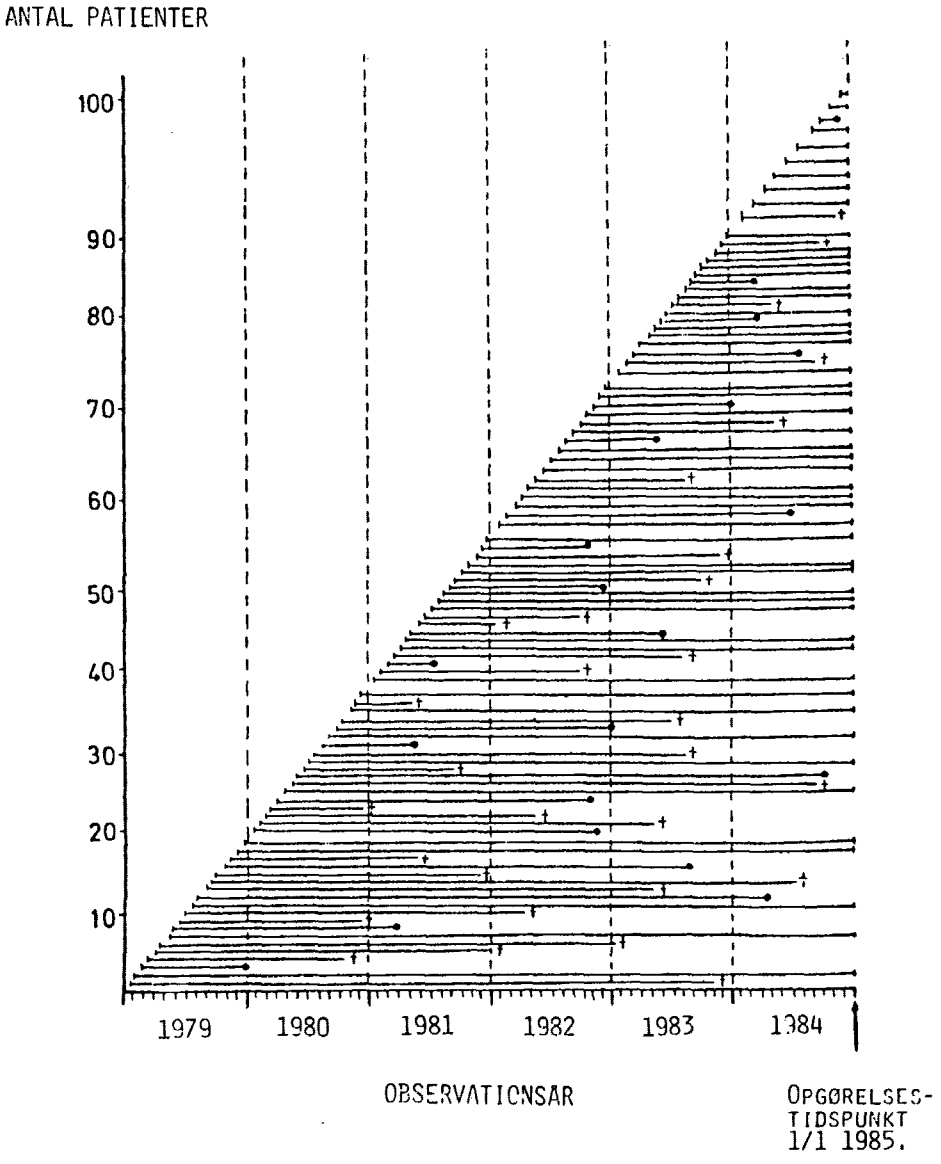
Imidlertid er det først i midten af dette århundrede, at en egentlig målrettet forskning af modeller for overlevelsesdata er vokset frem, og drivkraften bag udviklingen er primært sket i samarbejde mellem statistikere på den ene side og læger samt biologer på den anden. Det skal også nævnes, at forsikringsmatematikere - aktuarer - som indlysende har vitale interesser i at vide noget om enkeltpersoners risiko for at pådrage sig sygdom eller dø, også har været med til at drive forskningsarbejdet fremad.

b. Et eksempel på en undersøgelse.

Allerede når man begynder på overvejelserne om at lave en overlevelsesanalyse, må man være opmærksom på de vanskeligheder, der kan optræde. En væsentlig ting, som man bør holde sig for øje, er, hvor lang tid dataindsamlingen skal foregå, og hvilke kriterier, der er for at en person bliver registreret som udgået af undersøgelsen. I de fleste tilfælde vil afslutningen være død. En undersøgelse kunne gå ud på at undersøge,

hvor lang tid, der vil gå indtil kræftpatienter døde. For at få nok personer med i undersøgelsen ville man inddrage alle personer, som fik konstateret kræft i årene 1979 til 1985. De personer, der er med i undersøgelsen, kunne så i løbet af undersøgelsen være døde af kræft eller gået ud af undersøgelsen af en for undersøgelsen uvæsentlig grund, eller de kunne være i live, når undersøgelsen slutter; altså undersøgelsens opgørelsestidspunkt.

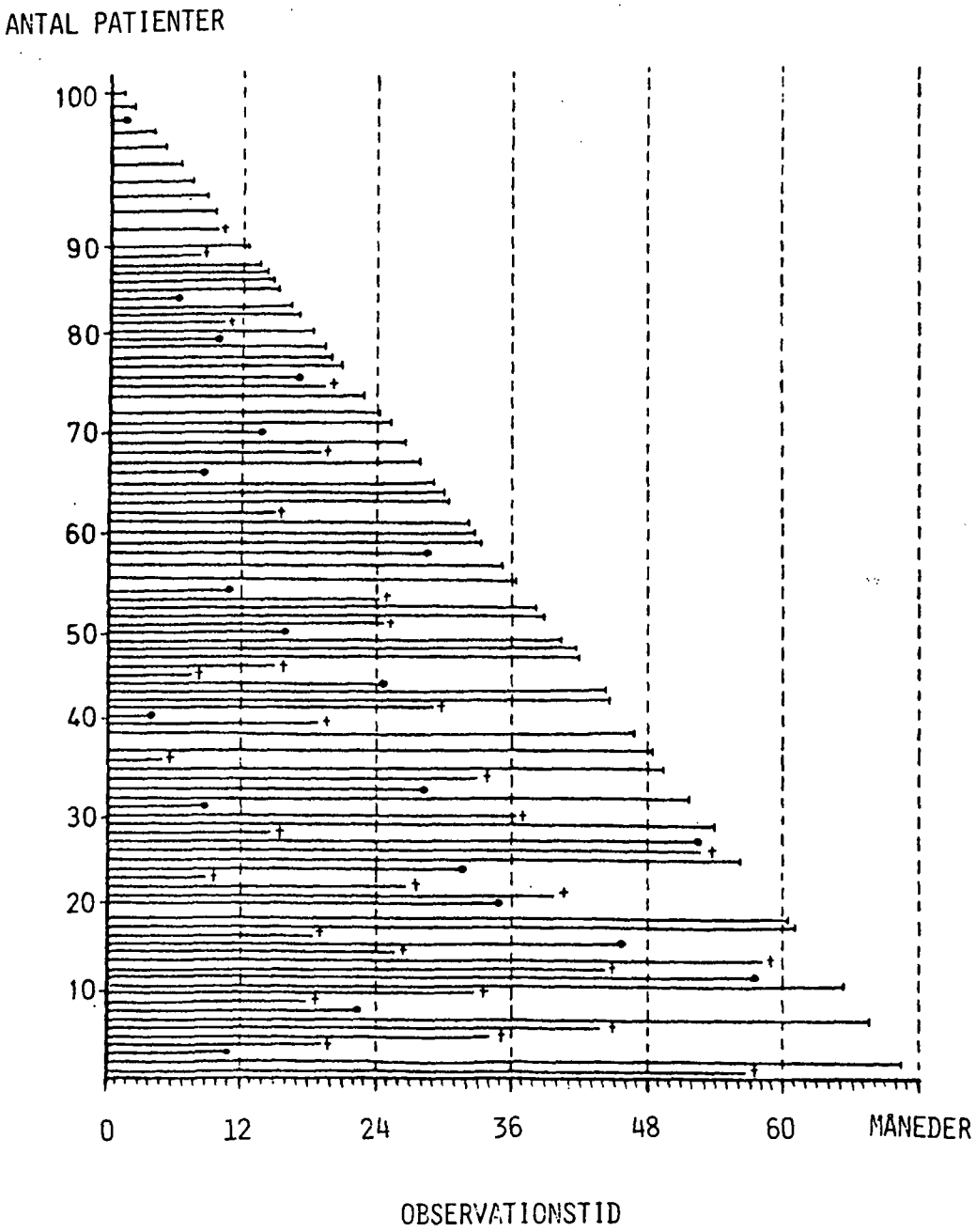
Figur 2.2 Levetider for kræftpatienter, fiktivt materiale.



Hver patient er repræsenteret ved en streg, der angiver observationstiden indtil:
| opgørelsestidspunktet.
+ tidspunkt for død.
• afgang af uvæsentlig grund.

Hvis man antager, at det ikke har nogen indflydelse på overlevelsestiden, hvornår man kom ind i undersøgelsen, vil det være hensigtsmæssigt at rykke alle pindene hen så de starter til samme tidspunkt kaldet t_o.

Figur 2.3 Ordnet levetider for kræftpatienter, fiktivt materiale.



På figur 2.3 hvor alle patienter er ordnet på denne måde, er der så tre mulige slutobservationer: + død, | i live på opgørelsestidspunktet og • gået ud af en for undersøgelsen uvæsentlig årsag.

De to sidste statusgrupper kaldes censurerede, umiddelbart kan vi måske ikke bruge dem til noget; men hvis vi tænker os om, ved vi jo, at de personer har overlevet mindst indtil deres opgørelsestidspunkt, hvilket er væsentlig, hvis man vil undersøge, hvor stor sandsynligheden er for at have overlevet mere end to år. I dette tilfælde har 53 overlevet de første to år; hvis vi nu havde undladt at bruge oplysningerne fra de censurerede ville kun 15 have overlevet. Det er altså en stor gevinst at inddrage de censureredes oplysninger i vores overlevelsesdatamateriale. Betingelsen for at bruge censureringerne er, at de falder jævnt i hele undersøgelsesperioden. De censureringer, der opstår når undersøgelsen slutter kalder vi for nemheds skyld tidscensureringer.

c. Basale begreber.

I en homogen population (dvs. i en population, hvor alle tænkes at have samme dødelighed), kan levetidernes variation beskrives ved en overlevelsesfordeling - og i dette afsnit præsenteres fire begreber, der entydigt definerer overlevelsesfordelingen. Disse fire begreber er: tæthedsfunktionen $f(t)$, overlevelsesfunktionen $S(t)$, dødsintensiteten $\lambda(t)$, og den integrerede (kumulerede) intensitet $\Lambda(t)$. Hvad dækker begreberne så over?

Tæthedsfunktionen $f(t)$. Sandsynligheden for at et individ dør i løbet af et lille interval $[t;t+\Delta t[$ er $f(t)\cdot\Delta t$.

Overlevelsesfunktionen $S(t)$. $S(t)$ beskriver sandsynligheden for at være i live til tiden t. $S(t)$ er en

aftagende funktion, hvor $S(0)=1$ og $S(\infty)=0$. Overlevelsesfordelinger beskrives også ved en fordelingsfunktion $F(t)$, hvor sammenhængen med $S(t)$ ganske enkelt er $S(t) = 1-F(t)$.

Dødsintensiteten $\lambda(t)$. Sandsynligheden for at et individ dør i intervallet $[t;t+\Delta t[$ forudsat, det var i live til tiden t, er $\lambda(t)\cdot\Delta t$.

Den integrerede (kumulerede) intensitet $\Lambda(t)$. $\Lambda(t)$ defineres som integralet over $\lambda(t)$. $\Lambda(t)=\int_0^t \lambda(u)du$.

Hvis blot man kender én af de fire ovennævnte funktioner, kan man beregne de tre øvrige. I det følgende vil denne påstand verificeres.

Sammenhængen mellem $f(t)$ og $S(t)$ kan forklares på denne måde: Hvis et individ dør efter tid t, må det dø enten i det første interval efter t eller i det andet eller i det tredie eller i det fjerde eller.... Det betyder, at sandsynligheden for at være i live til tid t er summen af sansynlighederne for at dø i et af intervallerne efter tid t. Det skrives således:

$$S(t) \approx f(t)\Delta t + f(t+\Delta t)\Delta t + f(t+2\Delta t)\Delta t + ..$$

For Δt gående mod 0 fås

$$(1) \qquad S(t) = \int_t^\infty f(u)du.$$

Omvendt kan man også udtrykke $f(t)$ ved $S(t)$.

$$S'(t) = [f(u)]_t^\infty$$

$$S'(t) = f(\infty) - f(t)$$

men da $f(\infty) = 0$ bliver udtrykket

$$(2) \qquad f(t) = - S'(t).$$

Sammenhængen mellem $\lambda(t)$, $S(t)$ og $f(t)$ fås ved denne

argumentation: Sandsynligheden for at dø i intervallet $[t;t+\Delta t[$, som er $f(t)\cdot\Delta t$, er lig sandsynligheden for at være i live til t, altså S(t) multipliceret med sandsynligheden for at dø i intervallet $[t;t+\Delta t[$ givet man har været i live til tid t, altså $\lambda(t)\cdot\Delta t$. Dette giver

$$f(t)\cdot\Delta t = S(t)\cdot\lambda(t)\cdot\Delta t$$

$$\lambda(t) = \frac{f(t)}{S(t)}$$

hvis $-S'(t)$ indsættes i stedet for $f(t)$

$$\lambda(t) = -\frac{d}{dt}(S(t))\cdot\frac{1}{S(t)}$$

og videre ved hjælp af differentiationsregneregler

$$\lambda(t) = -\frac{d}{dt}(\ln(S(t))). \tag{3}$$

Sammenhængen mellem $\Lambda(t)$ og S(t) fås således:

$$\Lambda(t) = \int_0^t \lambda(u)du$$

Indsættes det foregående (3) resultat fås

$$\Lambda(t) = \int_0^t \left(-\frac{d}{dt}(\ln S(u))\right) du$$

$$\Lambda(t) = -[\ln S(t)]_0^t$$

$$\Lambda(t) = -\ln S(t).$$

Dette udtryk kan omformes til

$$S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u)du\right). \tag{4}$$

Som det ses, vil man med kendskab til blot en af de fire funktioner automatisk også kende de tre øvrige. Med denne relative grundige gennemgang af begreberne skal deres substantielle betydning for overlevelses-modeller understreges.

Til slut et lille eksempel på anvendelser af formlerne.

Hvis $\lambda(t) = k$, altså konstant dødsintensitet over tiden, får man $\Lambda(t)$ på følgende måde

$$\int_0^t \lambda(u)du = \int_0^t k\ du$$

$$\Lambda(t) = k \cdot t,$$

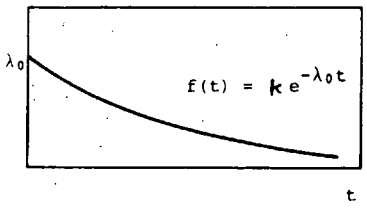
med dette udtryk kan vi opstille formlerne for både overlevelsesfunktionen og tæthedsfunktionen.

$$S(t) = \exp(-k \cdot t)$$

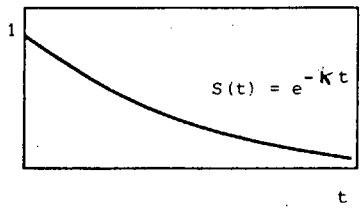
$$f(t) = k \cdot \exp(-k \cdot t).$$

På nedenstående figur er kurverne for funktionerne indtegnet.

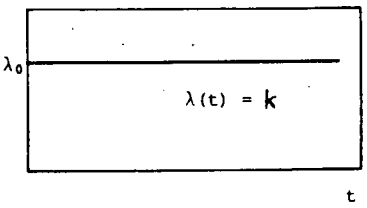
Figur 2.4. Overlevelseskurver.



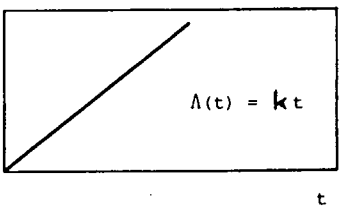
Tæthedsfunktionen.



Overlevelsesfunktionen.



Dødsintensiteten.



Den integrerede intensitet.

Vi må straks sige, at det er yderst sjældent at man kan postulere, at der er konstant dødsintensitet; og det optræder ikke i Cox-modellen, som taler om proportionalitet mellem dødsintensiteterne.

d. Simple estimater.

De to estimater viher vil gennemgå er Kaplan-Meier estimatet for overlevelsesfunktionen $S(t)$ og Nelson estimatet for den integrerede intensitet $\Lambda(t)$. At det er estimater, betyder at det som f.eks. Kaplan-Meier estimatet viser, ikke er den rigtige overlevelsesfunktion $S(t)$, men et skøn over denne, som skrives $\hat{S}(t)$ og ligeledes for Nelson estimatet $\hat{\Lambda}(t)$.

Som et eksempel på et estimat (skøn) over $S(t)$ kunne man indtegne overlevelsesfunktionen for Danmarks befolkning ved hjælp af en dødelighedstavle fra statistisk årbog. På nedenstående figur er der et uddrag af en dødelighedstavle.

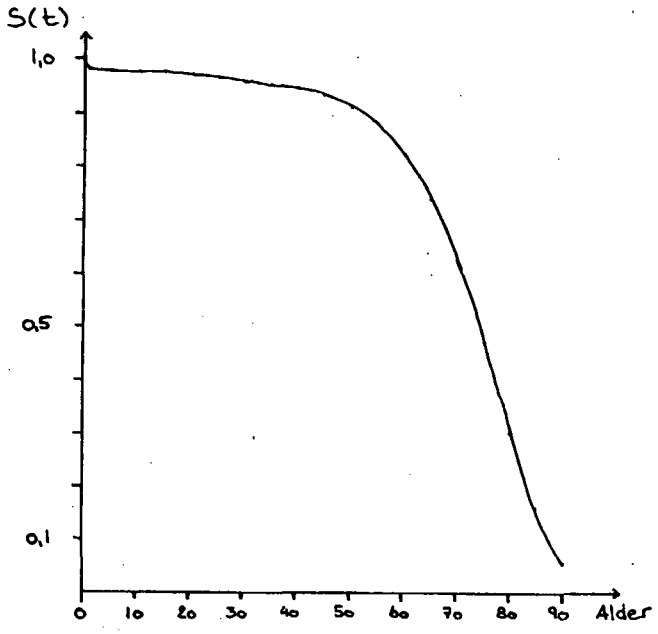
Figur 2.5 Udklip fra en dødelighedstavle.

	Aldersklassens dødshyppighed	Overlevende		Aldersklassens dødshyppighed	Overlevende		Aldersklassens dødshyppighed	Overlevende
0 År ...	1 924	100 000						
1 » ...	115	98 076	31 » ...	122	95 753	61 » ...	1 947	80 647
2 » ...	87	97 963	32 » ...	123	95 637	62 » ...	2 130	79 077
3 » ...	74	97 878	33 » ...	124	95 519	63 » ...	2 329	77 393
4 » ...	65	97 806	34 » ...	141	95 401	64 » ...	2 601	75 590
5 » ...	57	97 742	35 » ...	157	95 266	65 » ...	2 869	73 624
6 » ...	55	97 687	36 » ...	158	95 116	66 » ...	3 115	71 511
7 » ...	55	97 633	37 » ...	170	94 966	67 » ...	3 332	69 284
8 » ...	54	97 579	38 » ...	189	94 805	68 » ...	3 659	66 975
9 » ...	51	97 526	39 » ...	217	94 625	69 » ...	4 097	64 525
10 » ...	48	97 477	40 » ...	237	94 420	70 » ...	4 533	61 881
11 » ...	39	97 430	41 » ...	253	94 196	71 » ...	4 886	59 076
12 » ...	34	97 391	42 » ...	279	93 958	72 » ...	5 250	56 189
13 » ...	40	97 358	43 » ...	293	93 696	73 » ...	5 689	53 240
14 » ...	49	97 319	44 » ...	331	93 421	74 » ...	6 167	50 211
15 » ...	61	97 271	45 » ...	377	93 112	75 » ...	6 578	47 114
16 » ...	73	97 212	46 År ...	419	92 761	76 » ...	7 112	44 015
17 » ...	91	97 140	47 » ...	451	92 372	77 » ...	7 888	40 885
18 » ...	111	97 052	48 » ...	497	91 956	78 » ...	8 605	37 660
19 » ...	117	96 944	49 » ...	554	91 499	79 » ...	9 370	34 419
20 » ...	116	96 831	50 » ...	595	90 991	80 » ...	10 248	31 194
21 » ...	107	96 719	51 » ...	673	90 450	81 » ...	11 198	27 997
22 » ...	97	96 615	52 » ...	755	89 841	82 » ...	12 431	24 862
23 » ...	96	96 521	53 » ...	838	89 162	83 » ...	13 335	21 771
24 » ...	96	96 429	54 » ...	934	88 415	84 » ...	14 302	18 868
25 » ...	92	96 337	55 » ...	1 014	87 589	85 » ...	15 744	16 170
26 » ...	89	96 248	56 » ...	1 143	86 701	86 » ...	17 314	13 624
27 » ...	97	96 162	57 » ...	1 293	85 710	87 » ...	19 193	11 265
28 » ...	104	96 069	58 » ...	1 423	84 602	88 » ...	20 567	9 103
29 » ...	109	95 969	59 » ...	1 579	83 398	89 » ...	21 678	7 231
30 » ...	115	95 864	60 » ...	1 747	82 081	90 » ...	23 427	5 663

Dødelighedstavle, beregnet på grundlag af erfaringerne i årene 1966-1970.

Udfra antallet af overlevende fra de enkelte år er det muligt at lave en graf over overlevelsesfunktionen, når udgangspunktet er 100.000 mennesker og at $S(0) = 1$. På figuren nedenfor er kurven for overlevelsesfunktionen indtegnet.

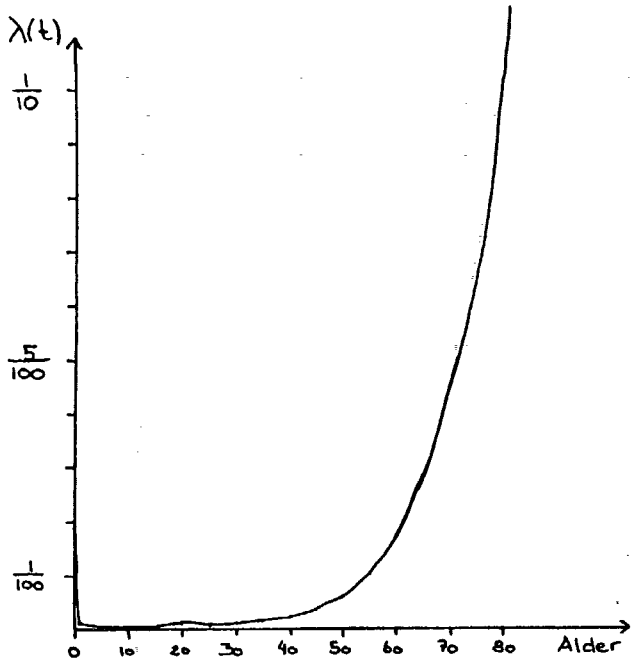
Figur 2.6 Overlevelsesfunktion for mænd.



Udfra oplysningerne kan man også beregne $\lambda(t)$ dødsintensiteten. I kolonnen alderklassens dødshyppighed står tallet for hvor mange, der dør til hver tid udaf 100.000 individer. Men det er jo det sammen som $\lambda(t)\Delta(t)$ fordi det netop beskrev sandsynligheden for at dø i et interval; forudsat at man i live ved intervallets start. På næste side er grafen for dødsintensiteten indtegnet.

(At tegne disse kurver, som om de var differentiable er selvfølgelig en tilsnigelse; idet vi ikke på streng matematisk vis kan differentiere dem. Men alligevel er de lavet udfra daglige observationer, og må derfor betegnes som et usædvanligt fint skøn over den "virkelige" fordeling. Det berettiger til at tegne den pænde, bløde kurve.)

Figur 2.7 Dødsintensiteten for mænd.



De ting, som vi her har regnet ud, er ikke beregnet efter en regneforskrift, men er bare aflæst. Hvis ens data ikke var en dødelighedstavle, men en tabel over hvilke hændelser der skete til bestemte tidspunkter, bliver man nødt til at bruge et Kaplan-Meier estimat til at beregne S(t). På figuren på næste side er dødstiderne fra pindediagrammet (fig. 2.3) opskrevet med et tidsinterval på en måned.

I disse tal indgår de omtalte censureringer, altså personer vi ikke ved hvornår er døde; men kun at de var i live til en bestemt tid. På grund af censureringerne kan man ikke bare som før gå igang med at indtegne overlevelsesfunktionen, man er nødt til at bruge et Kaplan-Meier estimat for at få et skøn over S(t). Kaplan-Meier estimatoren ser således ud:

$$(5) \qquad \hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{m_i}{Y(t_i)}\right)$$

hvor m_i angiver antallet af døde til tiden t_i og $Y(t_i)$ angiver antallet af personer i live lige inden tiden t_i. Over alle dødstidspunkterne skal udtrykket multipliceres.

Figur 2.8 Skema over dødstidspunkterne fra figur 2.3.

Dødstidspunkt	I live	Antal døde
5	95	1
7	91	1
8	89	2
9	84	1
10	81	1
15	73	3
18	65	2
19	62	4
24	53	1
25	50	1
26	48	2
29	42	1
32	37	1
33	34	1
36	29	1
40	25	1
43	21	2
52	12	1
56	9	1
58	6	1

Med udtrykket fra (5) kan vi så beregne $\hat{S}(t)$ for tallene fra figur 2.8. $\hat{S}(9)$ beregnes på følgende måde:

$$\hat{S}(9) = (1 - \frac{1}{95}) \cdot (1 - \frac{1}{91}) \cdot (1 - \frac{2}{89}) \cdot (1 - \frac{1}{84})$$

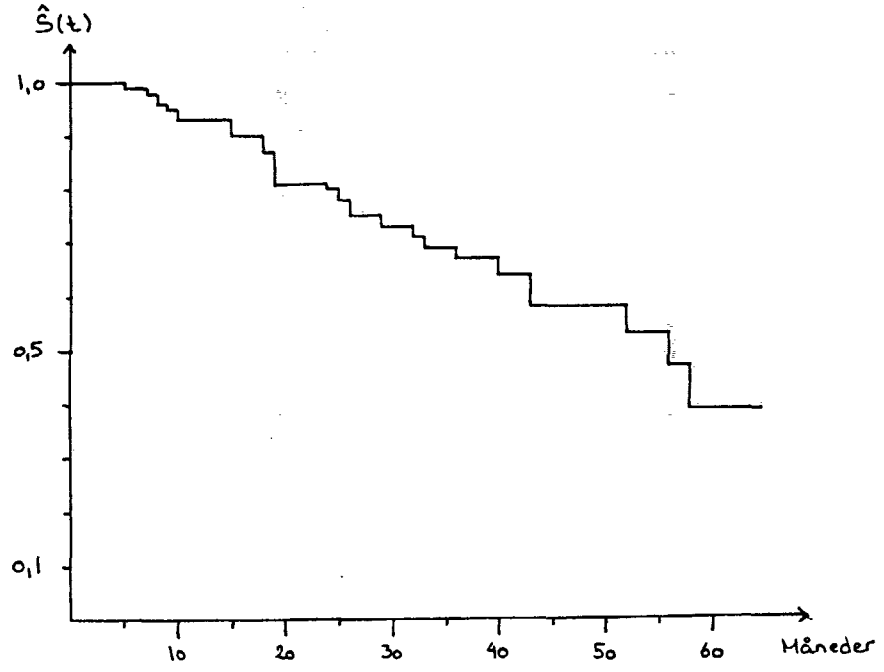
$$\hat{S}(9) = 0{,}9452.$$

Overlevelsesfunktionen til tid t kan altså estimeres ved produktet af en række sandsynligheder, der hver især er fremkommet på grundlag af levetider der er mindre end t.

Kaplan-Meier estimatet er en aftagende, stykkevis konstant funktion, "springpunkterne" svarer til de observerede dødstidspunkter.

På næste figur er Kaplan-Meier estimatet for det fiktive patientmateriale indtegnet, og bemærk i forhold til grafen fra figur 2.6 at funktionen ikke er differentiabel.

Figur 2.9 Kaplan-Meier estimat for figur 2.8.



Ud over Kaplan-Meier estimatet er Nelson estimatet meget anvendt, det har følgende udseende:

$$(6) \qquad \hat{\Lambda}(t) = \sum_{t \le t_i} \frac{m_i}{Y(t_i)}.$$

Den integrerede intensitet estimeres ved summen af sandsynligheder, der fremkommer på baggrund af de dødstidspunkter, der er mindre eller lig med t.

Nelson estimatet er en voksende, stykkevis konstant funktion, hvor "springpunkterne" svarer til de observerede dødstidspunkter. Nelson estimatet har ikke nogen egentlig tolkning, men er meget anvendeligt især til de vigtige proportionalitets test, som vi vender tilbage til i kapitel 3 om Cox-modellen.

e. Ikke parametriske test til sammenligning af flere gruppers overlevelsesfordeling.

Når man vil sammenligne to eller flere behandlingsgruppers overlevelsesfordeling, kan man sammenligne

Kaplan-Meier estimaterne for grupperne. Dette gøres grafisk ved at tegne Kaplan-Meier estimaterne for grupperne ind i samme diagram. Man kan vurdere om kurverne adskiller sig signifikant fra hinanden på et givet tidspunkt. Imidlertid er det indlysende, at en sammenligning, hvori hele det betragtede tidsforløb indgår, tydeligere fortæller om der er signifikant forskel eller ej grupperne imellem.

Dette problem imødegås ved hjælp af en række test, hvori man udnytter sit kendskab til hele overlevelsesfunktionens forløb. Disse test kan opdeles i ikke parametriske og parametriske tests. Ikke parametriske test forudsætter ikke en bestemt parametrisk form af overlevelsesfunktionen - dvs. man behøver ikke at kende det eksakte udtryk for overlevelsesfunktionen, "funktionens formel". Modsat gælder for parametriske test, at overlevelsesfunktionen er kendt; pånær et endeligt antal parametre.

En fordel ved ikke parametriske tests er, at de ikke forudsætter andet end at materialet ser ud, som vi hidtil har beskrevet det. Som vi skal se i det følgende eksempel med log rank testet, forudsætter disse tests intet kendskab til de enkelte personers egenskaber, og afvigelser fra forventede til observerede værdier udregnes kun udfra viden om befolkningens kvantitative tilstand; dvs. oplysninger af typen: Hvormange er i live, og hvornår.

Alligevel viser det sig i praksis, at nogle tests er mere følsomme end andre f.eks. overfor sammenfaldende dødstidspunkter; at nogle er bedre, hvis dødsintensiteterne mellem grupperne er proportionale (log rank testet) eller lignende.

Hvis man på den anden side har gjort sig nogle antagelser om sit materiale, har beskrevet dødeligheden ved hjælp af nogle parametre, vil det være naturligt at inddrage disse mere detaljerede oplysninger om den forventede dødelighed i testet. Vi vil senere i pro-

jektet støde på disse såkaldte parametriske tests, der bearbejder materialet for at få den bedste opbygning af Cox-modellen. Sådanne tests (f.eks. kvotienttest og scoretest) er mere forfinede end de ikke parametriske; men hviler altså på nogle flere forudsætninger

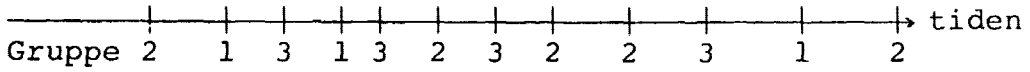
Der findes imidlertid flere forskellige tests til sammenligning af simple overlevelsesfordelinger; men når materialet rummer censureringer, falder der flere fra. Vi har valgt kun at fremdrage et test: Log rank testet. Dette test er især anvendeligt når de grupper der testes er proportionale. Det vil sige at

$$\lambda_2(t) = \theta \cdot \lambda_1(t) \qquad \text{for alle } t,$$

men også selvom dette ikke er tilfældet, kan man bruge log rank testet.

Testet findes i flere udgaver bl.a. hvor der er to grupper der testes og yderligere i to forskellige udgaver, når flere end to grupper sammenlignes. Det, som der i alle tilfælde bliver testet for, er om dødsintensiteten $\lambda(t)$ er ens i alle grupper. Man tester altså den såkaldte nulhypotese H_o. I så fald er log rank testet chi-i-anden fordelt med en eller flere frihedsgrader. I tilfælde, hvor der er flere end to grupper og de viser sig signifikant forskellige, kan man ikke vide hvilken (hvis der er netop en) gruppe, der adskiller sig fra de andre.

Log rank testet bruges altså til at sammenligne dødsintensiteten mellem forskellige grupper. For nemmere at vise hvordan det bruges, gennemgåes testet ved hjælp af et eksempel. Vi forestiller os, at vi har tre grupper, hvor vi vil teste, om de har samme dødsintensitet. I de tre grupper falder dødsfaldene således:



I testet interesserer vi os ikke for selve tidsforløbet, men blot rækkefølgen af dødsfaldene, da vi ikke hér studererer dødsintensiteterne, men interesserer os for forskellene mellem grupperne!
Log rank testet består af en rækkevektor, en covarians matrix og en søjlevektor. Produktet af disse giver ved udregning chi-i-anden størrelsen, med (antal grupper - 1) som antallet af frihedsgrader. Opstillingen ser således ud

$$(7) \qquad \underline{U}(0)' \cdot V^{-1} \cdot \underline{U}(0)$$

her refererer nullet (0) til nulhypotesen.

Da $\underline{U}(0)'$ blot er $\underline{U}(0)$ transponeret, er det nok at gennemgå $\underline{U}(0)$ og V.

$\underline{U}(0)$ beregnes som de observerede hændelser minus de forventede

$$\underline{U}(0) = \underline{O} - \underline{E}$$

$\underline{O}$ er det observerede antal af døde i grupperne.

$\underline{E}$ beregnes efter følgende formel

$$(8) \qquad \underline{E} = \sum_{i=1}^{k} \left(\frac{m_i}{Y(t_i)} \cdot \sum_{l \in R(t_i)} \underline{z}_l \right),$$

hvor der summeres over de k dødsfald.

m_i = antallet af døde til tid t_i.
$Y(t_i)$ = antallet af personer i live lige inden t_i.
$\underline{z}_l$ = antal personer i hver af grupperne lige inden tid t_i (risikomængden).

Som det ses af regneudtrykket, bruges tiden kun til at sætte gruppernes dødsfald i rækkefølge. Brugt på vores eksemplet beregnes $\underline{E}$ på denne måde:

Ved første dødsfald dør der 1 person $m_i = 1$. Risikomængden i $\underline{z}$ er henholdsvis 3, 5 og 4. Ved anden dødsfald dør der også 1, her er risokomængden i $\underline{z}$ 3, 4 og 4. På denne måde fortsætter man indtil alle dødsfaldene indgår i formeludtrykket. På næste side er hele udregning skrevet op.

$$\underline{E} = \frac{1}{12}\begin{pmatrix}3\\5\\4\end{pmatrix} + \frac{1}{11}\begin{pmatrix}3\\4\\4\end{pmatrix} + \frac{1}{10}\begin{pmatrix}2\\4\\4\end{pmatrix} + \frac{1}{9}\begin{pmatrix}2\\4\\3\end{pmatrix} + \frac{1}{8}\begin{pmatrix}1\\4\\3\end{pmatrix} +$$

$$\frac{1}{7}\begin{pmatrix}1\\4\\2\end{pmatrix} + \frac{1}{6}\begin{pmatrix}1\\3\\2\end{pmatrix} + \frac{1}{5}\begin{pmatrix}1\\3\\1\end{pmatrix} + \frac{1}{4}\begin{pmatrix}1\\2\\1\end{pmatrix} + \frac{1}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} +$$

$$\frac{1}{2}\begin{pmatrix}1\\1\\0\end{pmatrix} + \frac{1}{1}\begin{pmatrix}0\\1\\0\end{pmatrix} \qquad <=>$$

$$\underline{E} = \begin{pmatrix}2,663\\6,130\\3,208\end{pmatrix}$$

$$\underline{U}(0) = \begin{pmatrix}3\\5\\4\end{pmatrix} - \begin{pmatrix}2,663\\6,130\\3,208\end{pmatrix} = \begin{pmatrix}0,337\\-1,130\\0,792\end{pmatrix}$$

Matricen V beregnes udfra følgende generelle formel

$$(9) \qquad V_{hj} = \sum_{i=1}^{k} m_i \left(\frac{Y_j(t_i)}{Y(t_i)} \delta_{hj} - \frac{Y_h(t_i) \cdot Y_j(t_i)}{Y(t_i)^2} \right) \cdot \left(\frac{Y(t_i) - m_i}{Y(t_i) - 1} \right)$$

$Y_j(t_i)$ = antallet af personer i live lige inden tid t_i i den j'te gruppe.

$Y_h(t_i)$ = antallet af personer i live lige inden tid t_i i den h'te gruppe.

δ_{hj} = en funktion der er 0 når $h \neq j$ ellers 1.

Udtrykket bliver nemt at regne med, hvis der kun dør en person til hvert tidspunkt. Derudover kan det deles op i to, alt efter om h = j eller ej.

Hvis h = j.

$$(10) \qquad V_{hj} = \sum_{i=1}^{k} \frac{Y_h(t_i)}{Y(t_i)} \cdot \left(1 - \frac{Y_h(t_i)}{Y(t_i)} \right)$$

Hvis h ≠ j.

$$(11) \qquad V_{hj} = \sum_{i=1}^{k} - \frac{Y_h(t_i) \cdot Y_j(t_i)}{Y(t_i)^2}$$

I vores eksempel hvis h = j og h = 1 dvs. vi arbejder med gruppe 1 bliver:

$$V_{11} = \frac{1}{12}(1 - \frac{1}{12}) + \frac{1}{11}(1 - \frac{1}{11}) + \frac{1}{10}(1 - \frac{1}{10}) + \frac{1}{9}(1 - \frac{1}{9}) + \frac{1}{8}(1 - \frac{1}{8}) + \frac{1}{7}(1 - \frac{1}{6}) + \frac{1}{5}(1 - \frac{1}{5}) + \frac{1}{4}(1 - \frac{1}{4}) + \frac{1}{3}(1 - \frac{1}{3}) + \frac{1}{2}(1 - \frac{1}{2}) \quad <=>$$

$$V_{11} = 1{,}909.$$

I eksemplet hvis h ≠ j og h = 1, j = 2. Her inddrages både gruppe 1 og 2. Nu bliver

$$V_{12} = -\frac{3 \cdot 5}{12^2} - \frac{3 \cdot 4}{11^2} - \frac{2 \cdot 4}{10^2} - \frac{2 \cdot 4}{9^2} - \frac{1 \cdot 4}{8^2} - \frac{1 \cdot 4}{7^2} - \frac{1 \cdot 3}{6^2} - \frac{1 \cdot 3}{5^2} - \frac{1 \cdot 2}{4^2} - \frac{1 \cdot 1}{3^2} - \frac{1 \cdot 1}{2^2} \quad <=>$$

$$V_{12} = -1{,}216.$$

Så snart man har fundet V_{12}, har man også fundet V_{21}, for som det kan ses i (11), er det kun faktorerne, der bliver byttet om på.

Hvis vi et øjeblik vender tilbage til $\underline{U}(0) = \underline{O} - \underline{E}$, kan det ses, at de observerede dødsfald sammenlagt (3+5+4) giver 12 ligesom de forventede (2,663+6,13+ 3,208) også giver 12. Det betyder, at det kun er nødvendigt at regne med (K-1) grupper i log rank testet, da en gruppe hele tiden er beskrevet ved de øvrige. For vores eksempel betyder det, at vi kun mangler at beregne V_{22}, som bliver 2,668.

Log rank teststørrelsen bliver så:

$$\chi_2^2 = (0{,}337;-1{,}13)\begin{pmatrix} 1{,}909 & -1{,}216 \\ -1{,}216 & 2{,}668 \end{pmatrix}^{-1}\begin{pmatrix} 0{,}337 \\ -1{,}130 \end{pmatrix}$$

$$\chi_2^2 = 0{,}502.$$

I forhold til et signifikans niveau på 5% kan vi ikke forkaste hypotesen, altså er dødsintensiteten ens i de tre grupper.

Det her gennemregnede eksempel indeholder hverken censureringer eller flere dødsfald samtidig, men udregnings proceduren er den samme i disse tilfælde, blot skal man huske at bruge formel (9).

For at give en forsmag på hvad der kan testes , kan vi her rævne nogle få ting, som også får betydning for den endelige Cox-model. En væsentlig antagelse ved Cox-modellen er at der er proportionale dødsintensiteter mellem alle personer. Det er derfor oplagt at teste om dette faktisk er tilfældet. I samme åndedræt kunne man så finde en eventuel overdødelighed i nogle af de grupper, der testes.

3. BESKRIVELSE AF COX-MODELLEN.

I modellen, der her skal præsenteres, antager man grundlæggende, at dødsintensisteterne for samtlige patienter er proportionale. Ideen med modellen er at kunne operere med mange baggrundsvariable. En baggrundsvariabel kan siges at være en bestemt oplysning, der er karakteristisk for et enkelt individ. Eksempelvis kan nævnes alder og køn. Man kan altså ikke uden videre sammenligne forskellige individer - man må først korrigere for baggrundsvariablernes indflydelse påde enkelte individers tilstand. Til forskel fra tidligere hvor vi sammenlignede homogene grupper, sammenligner vi nu enkelt individer - hvert individ udgør så at sige en gruppe.

Som beskrevet under ikke-parametriske metoder gjaldt følgende under antagelse om proportionale dødsintensiteter:

$$\lambda_2(t) = \theta\lambda_1(t) \quad \text{for alle } t$$

θ er proportionalitetsfaktoren eller overdødelighedsparametren og afhænger ikke af t. Hvis θ erstattes af e^{β} fås:

$$\lambda_2(t) = e^{\beta}\lambda_1(t)$$

I det følgende vil ideen i Cox-modellen vises udfra en sammenligning af overlevelsen i to patientgrupper.

Lad dødsintensisteten i gruppe 1 være $\lambda_0(t)$. Dødsintensiteten i gruppe 2 vil da være $\lambda_0(t)\cdot e^{\beta}$. Et givet individ - f.eks. individ nr i - vil naturligvis tilhøre enten gruppe 1 eller 2. Ved at indføre en baggrundsvariabel z_i kan man afgøre, hvilken af grupperne individet tilhører. z_i defineres som

$$z_i = \begin{cases} 0 & \text{hvis individet er i gruppe 1} \\ 1 & \text{hvis individet er i gruppe 2} \end{cases}$$

Dødsintensiteten for det i'te individ er da:

(1) $$\lambda_i(t) = \lambda_0(t) \cdot e^{\beta \cdot z_i}$$

Opgaven er nu at estimere β-regressionsparameteren og $\lambda_0(t)$ - den underliggende intensistet. Dette kan gøres ved hjælp af den såkaldte likelihood-funktion. Inden der fortsættes, vil denne kort præsenteres.

Antag man observerer n uafhængige variable $T_1 \ldots T_n$ med samme tæthedsfunktion $f(t,\theta)$, hvor θ er en ukendt parameter. Opgaven er at estimere/skønne over værdien θ, der tillægger observationerne størst mulig sandsynlighed.

Sandsynligheden for at T_1 falder i intervallet $[t_1, t_1+\Delta t_1]$ vil da være $f(t_1,\theta)\Delta t_1$ i følge den tidligere beskrevne definition af tæthedsfunktionen. Sandsynligheden for T_2 ligger i intervallet $[t_2, t_2+\Delta t_2]$ er $f(t_2,\theta)\Delta t_2$, osv. Da observationerne forudsættes uafhængige, vil sandsynligheden for, at observationerne netop falder som beskrevet, fås af produktet:

$$f(t_i,\theta)\Delta t_1 \cdot f(t_2,\theta)\Delta t_2 \ldots f(t_n,\theta)\Delta t_n = \prod_{i=1}^{n} f(t_i,\theta)\cdot\Delta t_i$$

Produktet $\prod_{i=1}^{n} f(t_i,\theta)$ kaldes likelihood-funktionen og benævnes $L(\theta)$.

Maximum-likelihood estimatoren $\hat{\theta}$ er den værdi for θ, der vil gøre $L(\theta)$ største mulig - eller ensbetydende hermed - den værdi, der gør $\ln L(\theta)$ størst mulig.

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(t_i,\theta).$$

Man kan finde $\hat{\theta}$ ved at finde maximumspunktet for $l(\theta)$. Man differentierer $l(\theta)$ med hensyn til θ og finder nulpunkt for den afledede $l'(\theta)$. For at forvisse sig om man faktisk har et toppunkt, må man finde den anden afledede - hvis denne er mindre end 0, dvs $l''(\theta)<0$, for alle θ, er man sikker.

Lad os nu betragte en patientgruppe, hvor der foruden dødsfald også observeres censureringer. Man har n observationen, hvoraf d af disse er dødsfald og

altså n-d er censureringer. Efter samme princip som lige beskrevet, vil de d dødsfald bidrage til likelihood-funktionen med $\prod_{i=1}^{d} f(t_i,\theta)$. Da man desuden har n-d censureringer - nemlig til tiderne $t^*_{d+1}, \ldots, t^*_n$ vil disse også yde et bidrag til likelihood funktionen. Om den første censurerede person vides blot at denne har overlevet til tid t^*_{d+1}, hvilket beskrives ved overlevelsesfunktionen $S(t^*_{d+1})$, den næste censurerede person har overlevet til t^*_{d+2} - man får $S(t^*_{d+2})$, osv. Da de censurerede observationer - i lighed med dødsfaldene - sker uafhængigt af hinanden - vil førstnævntes bidrag til likelihood funktionen være $\prod_{i=d+1}^{n} S(t_i)$.

Alt i alt fås likelihood funktionen til

$$(2) \qquad L(\theta) = \prod_{i=1}^{d} f(t_i,\theta) \cdot \prod_{i=d+1}^{n} S(t_i,\theta)$$

Likelihood funktionen udtrykker sandsynligheden for at det egentlige begivenhedsforløb faktisk skulle finde sted. $L(\theta)$ vil, hvis vi omregnede $f(t,\beta)$ og $S(t,\beta)$ og indsatte i (2), blive en kompliceret funktion af tiden og parametrene, hvor tiden udelukkende indgår i den underliggende dødsintensitet:

$$L(t,\beta) = \prod_{i=1}^{d} e^{\beta z_i} \lambda_0(t_i) \cdot \prod_{i=1}^{n} \exp(-e^{\beta z_i} \cdot \Lambda_0(t_i))$$

Hvis man ser bort fra tiden; men blot studerede rækkefølgen af begivenhederne forsimpler vi udtrykket en hel del.

Cox's likelihood funktion opererer kun med den parametriske del af $L(t,\beta)$, hvor hver begivenhed beskrives ved udtrykket

$$\frac{\lambda_i(t)}{\sum_{j \in R(t)} \lambda_j(t)}$$

$R(t)$ betegner risikomængden af patienter j, der har overlevet til tiden t - dvs de patienter, der er i

live og under observation lige inden tiden t. Med definitionen i afsnittet om Kaplan-Meier og Nelson estimater af Y(t), vil mængden R(t) bestå af Y(t) patienter.

Dette udtryk giver sandsynligheden for at når én nu skal dø ved næste begivenhed, at så er det person i med intensiteten $\lambda_i(t)$, der dør udaf en aktuel befolkning, hvis intensiteter er summeret i $\sum_R \lambda_j(t)$.

Når vi udskifter $\lambda_i(t)$ og $\lambda_j(t)$ med udtrykket fra (1) får vi beskrevet personernes dødsintensiteter v.hj.a. β og $\lambda_0(t)$:

$$L_c(\beta) = \prod_{i=1}^{d} \frac{\lambda_0(t_i)\, e^{\beta z_i}}{\sum_{j \in R(t_i)} \lambda_0(t_i) e^{\beta z_j}}$$

Vi kan nu dividere $\lambda_0(t)$ væk og får:

$$(3) \qquad L_c(\beta) = \prod_{i=1}^{d} \frac{e^{\beta z_i}}{\sum_{j \in R(t_i)} e^{\beta z_i}}$$

Hvis man følger proceduren som tidligere beskrevet findes $\hat{\beta}$ ved "at tage" ln til $L_c(\beta)$ og derpå differentiere:

$$l(\beta) = \ln L_c(\beta) = \sum_{i=1}^{d} \beta z_i - \sum_{i=1}^{d} \ln\Big(\sum_{j \in R(t_i)} e^{\beta z_j} \Big)$$

$$l'(\beta) = \sum_{i=1}^{d} z_i - \sum_{i=1}^{d} \frac{\sum_{j \in R(t_i)} e^{\beta z_j} \cdot z_j}{\sum_{j \in R(t_i)} e^{\beta z_j}}$$

For $l'(\beta) = 0$ fås

$$\sum_{i=1}^{d} z_i = \sum_{i=1}^{d} \frac{\sum_{j \in R(t_i)} e^{\beta z_j} \cdot z_j}{\sum_{j \in R(t_i)} e^{\beta z_j}}$$

Venstre side består af en sum af 0'er og 1-taller afhængig af, om personen, der dør, tilhører hhv. gruppe 1 eller gruppe 2. Med andre ord vil man få det observerede antal døde personer i gruppe 2. Lad dette være O_2.

Tælleren på højre side fortæller, hvis den j'te person i risikomængden tilhører gruppe 1 fås bidraget 0 ($0 \cdot e^{\beta 0}$) og hvis person j tilhører gruppe 2 fås bidraget e^{β} ($1 \cdot e^{\beta \cdot 1}$). Summeres fra i=1 til d vil man få antallet af levende personer i gruppe 2 "ganget med" e^{β}. I nævneren fås for person j tilhørende gruppe 1 bidraget $1 = e^{\beta \cdot 0}$ og tilsvarende for person j tilhørende gruppe 2 fås bidraget $e^{\beta} = e^{\beta \cdot 1}$. Summeres fra i=1 til d giver det antallet levende i gruppe 1 "plus" antallet af levende i gruppe 2 "ganget med" e^{β}.

For overblikkets skyld kan ligningen i prosa skrives således:

$$l'(\beta) = O_2 - \sum_{i=1}^{d} \frac{\text{antal levende i gr. 2} \cdot e^{\beta}}{\text{antal lev. i gr. 1 + lev. i gr.2} \cdot e^{\beta}}$$

(altsammen til dødstiderne t_i)

Man skal altså løse ligningen:

$$l'(\beta) = O_2 - \sum_{i=1}^{d} \frac{Y_2(t_i) \cdot e^{\beta}}{Y_1(t_i) + Y_2(t_i) \cdot e^{\beta}} = 0$$

Ved løsning af denne ligning må man benytte iterative metoder, som vi ikke nærmere vil komme ind på.

Opgaven er nu at sikre sig at $l''(\beta)<0$. Differentieres $l'(\beta)$ fås - idet O_2 er konstant:

$$l''(\beta) = - \sum_{i=1}^{d} \frac{(Y_1(t_i)+Y_2(t_i)e^{\beta}) \cdot Y_2(t_i)e^{\beta} \div Y_2(t_i)e^{\beta} \cdot Y_2(t_i)e^{\beta}}{(Y_1(t_i) + Y_2(t_i)e^{\beta})^2}$$

$$(4)\quad l''(\beta) = -\sum_{i=1}^{d} \frac{Y_1(t_i)\, Y_2(t_i)\, e^{\beta}}{(Y_1(t_i)+Y_2(t_i)e^{\beta})^2}$$

Det ses, at tælleren er positiv - idet $Y_1(t_i)$ og $Y_2(t_i)$ angiver et (positivt) antal. Dermed bliver hele udtrykket negativt - og man er sikker på, at for det fundne β giver $l(\beta)$ sin maximumsværdi.

Man er også interesseret i at estimere den underliggende intensitet $\lambda_0(t)$. Dette p.gr.a. at man dels vil kontrollere modellens forudsætninger om alle patienters dødsintensiteter er proportionale, og dels finde overlevelsesfunktionen. I stedet for at estimere $\lambda_0(t)$ vil man ofte estimere $\Lambda_0(t)$ - den integrerede underliggende intensitet. $\Lambda_0(t)$ kan estimeres udfra følgende

$$(5)\qquad \hat{\Lambda}_0(t) = \sum_{t_i \le t} \frac{m_i}{\sum_{j \in R(t_i)} e^{\hat{\beta} z_j}}$$

m_i: Antal dødsfald til t_i.

Det bemærkes, at hvis alle patienter tilhører gruppe 1 - da vil nævneren være antallet af patienter i live til t_i - eller $Y_1(t_i)$. Da vil $\Lambda_0(t)$ netop være Nelson estimatoren.

Udfra $\hat{\Lambda}_0(t)$ kan overlevelsesfunktionen estimeres udfra udtrykket:

$$(6)\qquad \hat{S}(t) = e^{-\hat{\Lambda}_0(t)}$$

Man kan udvide sine undersøgelser og sammenligne overlevelsen i K (K>2) patientgrupper og ikke blot i to grupper. Også hér antages proportionale dødsintensiteter og man antager desuden, at dødsintensiteten for behandlingsgruppe K er $\lambda_0(t)$ - den underliggende intensitet. En patient i gruppe 1 har dødsintensitet $\lambda_0 e^{\beta_1}$, en patient i gruppe 2 har dødsintensitet $\lambda_0 e^{\beta_2}$, osv. Nu indføres en række baggrundsvariable på føl-

gende vis:

$$z_{i1} = \begin{cases} 1 & \text{hvis patient i tilhører gr. 1} \\ 0 & \text{ellers} \end{cases}$$

$$z_{i2} = \begin{cases} 1 & \text{hvis patient i tilhører gr. 2} \\ 0 & \text{ellers} \end{cases}$$

$$z_{i,K-1} = \begin{cases} 1 & \text{hvis patient i tilhører gr. K-1} \\ 0 & \text{ellers} \end{cases}$$

For en tilfældig patient vil dødsintensiteten se således ud:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_{K-1} z_{i,K-1})$$

For en patient i gruppe K vil $z_{i1} = z_{i2} = \ldots = z_{i,K-1} = 0$

Man får således at $\exp(\beta_i)$ angiver forholdet mellem dødsintensiteten for en patient i gruppe i og en patient i gruppe K. $\exp(\beta_j)$ angiver forholdet mellem dødsintensiteten for en patient i gruppe j og en patient i gruppe K, osv. Forholdet mellem dødsintensiteten for en patient i gruppe i og en patient i gruppe j er $\exp(\beta_i - \beta_j)$.

Som tilfældet var ved sammenligning af to patientgrupper kan regressionsparametrene estimeres udfra Cox's likelihood funktion. Man får

$$(7) \qquad L_c(\beta) = \prod_{i=1}^{d} \frac{\exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_{K-1} z_{i,K-1})}{\sum_{j \in R(t_i)} \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_{K-1} z_{i,K-1})}$$

Som nævnt antages proportionale dødsintensiteter i Cox's regressionsmodel. Det må derfor være på sin plads at kontrollere, at antagelsen holder. Hvis der er tale om to grupper, skal der kontrolleres for følgende:

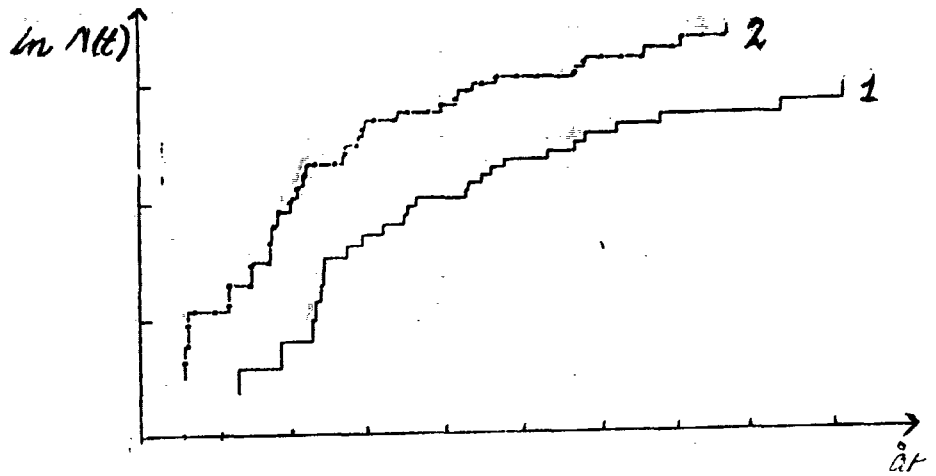
$$\lambda_2(t) = \lambda_1(t) e^{\beta} \qquad \text{for alle } t$$

Det kan udfra sammenhængen mellem de fire basale begreber, der er introduceret tidligere, omformuleres

til:

$$\Lambda_2(t) = \Lambda_1(t)e^{\beta} \Leftrightarrow \ln\Lambda_2(t) = \beta + \ln\Lambda_1(t)$$

Figur 3.1. ln $\Lambda_1(t)$ sammenlignet med ln $\Lambda_2(t)$ for et fiktivt materiale.



Afstanden mellem de to kurver skal være "rimelig" konstant, hvis der skal være basis for at antage proportionalitet. Afstanden vil være ca. lig med β.

I det foregående er der beskrevet en række baggrundsvariable - variable, der fortæller, hvilken behandlingsgruppe en vilkårlig person tilhører og altså er af typen 0 eller 1. Disse kaldes kvalitative baggrundsvariable. I det følgende beskrives en regressionsanalyse, hvori der indgår flere baggrundsvariable, som dels er kvalitative og dels er kvantitative (dvs. beskrevet med et tal, der henfører til en eller anden tings størrelse).

I vor undersøgelse om studietider på RUC har vi overvejet, hvilke forhold, der kunne påvirke et studiums længde. Blandt disse forhold - eller baggrundsvariable - har vi udvalgt to - alder ved studiestart og køn, som vi vil bruge i det efterfølgende lille eksempel.

Eksemplet skal illustrere en regressionsanalyse med flere baggrundsvariable og bygger på 1039 studenter, hvoraf 511 har fået eksamen og de resterende 528 er

censurerede. Man har to baggrundsvariable, z_{i1} og z_{i2}, som er henholdsvis kvalitativ og kvantitativ:

$$z_{i1} = \begin{cases} 1 & \text{hvis student i er en kvinde} \\ 0 & \text{hvis student i er en mand} \end{cases}$$

og

$$z_{i2} = \text{alder ved studiestart for student i}$$

Man kan nu opskrive følgende simple model ved at antage at dødsintensiteten for student i er givet ved:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 z_{i1} + \beta_2 z_{i2})$$

Straks må det understreges, at vi ikke kun tror, at alder ved studiestart og køn er de eneste baggrundsvariable, der har indflydelse på studietiderne, men vi har kun medtaget to for at forenkle og overskueliggøre problemstillingen.

Man kan nu estimere β_1 og β_2 udfra likelihood funktionen og får v.hj.a. en databehandling værdierne
$\lambda_i(t) = \lambda_0(t) \exp(0{,}0160\ z_{i1} - 0{,}0034\ z_{i2})$

Man kan tolke $\beta_1 = 0{,}0160$ således, at $\exp(0{,}0160)$ er forholdet mellem dødsintensiteterne for en mand og en kvinde med samme alder ved studiestart:

Dødsintensiteten for en vilkårlig kvinde med alder z_{i2}:
$$\lambda_0(t) \exp(0{,}0160) \exp(-0{,}0034\ z_{i2})$$
Dødsintensiteten for en vilkårlig mand med alder z_{i2}:
$$\lambda_0(t) \exp(0) \exp(-0{,}0034\ z_{i2})$$

Forholdet mellem dødsintensiteterne for en kvinde og en mand med samme alder er da: $\exp(0{,}0160)$.

På tilsvarende måde tolkes $\exp(-0{,}0034)$ som forholdet mellem dødsintensiteterne for to studenter af samme køn og en aldersforskel på 1 år. Hvis aldersforskellen er a år fås:

$$\lambda_{i+a}(t) = \lambda_0(t)\ \exp(\beta_1 z_{i1})\ \exp(-0{,}0034\ (x+a))$$

$$\lambda_i(t) = \lambda_0(t)\ \exp(\beta_2 z_{i1})\ \exp(-0{,}0034\ x)$$

$$\frac{\lambda_{x+a}(t)}{\lambda_i(t)} = \exp(-0{,}0034\ a)$$

Antagelsen for denne model er, at alder ved studiestart har samme påvirkning på overlevelsen for mænd såvel som kvinder, og påvirkningen, der hidrører fra kønnet, er den samme uanset alderen. Når dette gør sig gældende, antager man, at der <u>ingen vekselvirkning</u> er mellem køn og alder. For at undersøge denne antagelse nærmere kan man betragte en undersøgelse med tre baggrundsvariable:

$$z_{i1} = \begin{cases} 1 & \text{for student i er kvinde} \\ 0 & \text{for student i er mand} \end{cases}$$

$$z_{i2} = \text{alder for student i}$$

$$z_{i3} = z_{i1} \cdot z_{i2} = \begin{cases} \text{alder} & \text{for student i er kvinde} \\ 0 & \text{for student i er mand} \end{cases}$$

Derved fås følgende model:

$$\lambda_i(t) = \lambda_0(t)\ \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3})$$

Hvis antagelsen om ingen vekselvirkning skal holde vil $\beta_3 = 0$. I lighed med den tidligere model kan man også i denne estimere β_1, β_2 og β_3 ved likelihood funktionen.

<u>Model uden vekselv.</u>

$\hat{\beta}_1 = 0{,}0160$

$\hat{\beta}_2 = -0{,}0034$

$l_c(\hat{\beta}_1, \hat{\beta}_2) = -2928{,}7174$

<u>Model med vekselvirk.</u>

$\hat{\beta}_1 = -1{,}6325$

$\hat{\beta}_2 = -0{,}0258$

$\hat{\beta}_3 = 0{,}0728$

$l_c(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = -2923{,}7264$

Vi vil nu teste om $\beta_3 = 0$. Dette kan man gøre ved den såkaldte <u>kvotienttest</u>, som er defineret ved:

$$Q = \frac{L(\beta_i)}{L(\beta_j)}$$

hvor β_i er parametre fra den oprindelige model og β_j er parametre fra den nye model. I dette tilfælde svarerβ_i til (β_1,β_2) og β_j til $(\beta_1,\beta_2,\beta_3)$.

$2 \ln Q = 2|\ln (L(\beta_i)/L(\beta_j))| \Rightarrow$

$2 \ln Q = 2|l_c(\beta_1,\beta_2) - l_c(\beta_1,\beta_2,\beta_3)|$

Kvotienteststørrelsen for vort lille eksempel bliver $2 \ln Q = 9{,}9740$. Man kan sammenligne denne værdi med en χ^2-fordeling med 1 (dvs.3-2) frihedsgrad og et signifikansniveau på 10%. Den udregnede værdi er langt højere end tabelværdien 2,71, hvilket betyder, at der er al mulig grund til at tro, at vekselvirkningsparametren forbedrer modellen.

For en kvinde vil vi finde den alder, der gør kvindens dødsintensitet lig med den underliggende intensitet. For at dette skal opfyldes, skal eksponenten til eksponentialfunktionen være 0.

$$-1{,}6325\, z_{i1} - 0{,}0258\, z_{i2} + 0{,}0728\, z_{i3} = 0$$

Da $z_{i1} = 1$ (for kvinde) og $z_{i3} = z_{i1} \cdot z_{i2} \Leftrightarrow z_{i3} = z_{i2}$ fås

$$-1{,}6325 \cdot 1 - 0{,}0258\, z_{i2} + 0{,}0728\, z_{i2} = 0 \Leftrightarrow$$

$$\underline{z_{i2} = 34{,}7 \text{ år}}$$

Man får altså, at en kvinde på 34,7 år vil have en dødsintensitet lig den underliggende. Er en kvinde yngre end 34,7 år, vil eksponenten blive større end 0 og dødsintensiteten mindre end den underliggende. Eller med andre ord: Jo ældre en kvinde er ved start på et studie, desto hurtigere vil studiet kunne gennemføres <u>forudsat</u> alder og køn er de eneste baggrundsvariable.

Tilsvarende kan man for en mand finde den alder, der

gør dødsintensiteten lig den underliggende intensitet. Man får

$$-1,6325 \cdot 0 - 0,0258\ z_{i2} + 0,0728\ (0 \cdot z_{i2}) = 0$$

$$\underline{z_{i2} = 0.}$$

Vor "baggrundsperson" er her en mand på 0 år. En mand, der "har en alder" vil kunne gennemføre studiet hurtigere end vores (fiktive) baggrundsperson.

I alle følgende sammenhænge vil vi referere til en "baggrundsperson" som værende en (fiktiv) person med alle z-værdier = 0 samtidig.

Det skal i øvrigt bemærkes, at estimatet for kønseffekten hhv. effekten for alderen ved studiestart ændres i de to modeller. Med andre ord er der sammenhæng mellem de to variable, der indgår i modellen. Man siger, der er korrelation mellem variablene køn og alder, og der gælder, at jo mere vægt estimationsproceduren lægger på f.eks. køn, desto mindre vægt vil der lægges på den til køn relaterede variable alder. Med andre ord vil variablen alder "gemme" sig bag variablen køn.

Når man ved multipel regressionsanalyse - altså analyse af en model med flere baggrundsvariable - vil undersøge om en given baggrundsvariabels påvirkning er signifikant, må man gøre sig klart, at dette vil afhænge af de øvrige baggrundsvariable i modellen.

Problemet er så: På hvilken måde kan en given variabels effekt så testes. De to hyppigste måder er forward selection og backward elimination.

Princippet i den første er, at man for hver enkelt baggrundsvariabel, man vil have med i undersøgelsen, udregner den naturlige logaritme til likelihood funktionen og kvotientstørrelsen. Hvis man har n baggrundsvariable, vil man få n Cox-modeller af formen $\lambda_i(t) = \lambda_0(t) \exp(\beta_j z_j)$, hvor $j = 1, \ldots, n$. (1)

---

(1): Se næste side!

Hvor kvotienttesstørrelsen er mest signifikant, vælges den tilsvarende baggrundsvariabel ud. Denne model suppleres så med én af de tilbageværende n-1 variable - netop den, der nu er mest signifikant. Man får så en model af formen $\lambda_i(t)=\lambda_0(t)\exp(\beta_1 z_{1i}+\beta_2 z_{2i})$ osv. Således fortsætter man til man finder en variabel, der ikke giver signifikans, og en 'endelig' model er konstrueret, der i al korthed med vektor-symboler ser således ud:

(8) $$\lambda_i(t) = \lambda_0(t)\, e^{\boldsymbol{\beta}\, \mathbf{z}_i}$$

Ulempen ved denne metode er, at man - så længe ikke alle signifikante variable er inde i modellen - konkluderer udfra 'forkerte' modeller. En signifikant variabel vil jo ændre effekten på de variable, der var i modellen i forvejen.

Den anden metode - backward-elimination - er den 'modsatte' af ovennævnte metode, idet man hér starter med en model, hvor alle n baggrundsvariable er inkluderet. Den mindst signifikante vælges fra. I den næste etablerede model bestående af n-1 variable vælges atter den mindst signifikante fra og således fortsættes til man kun har variable, hvis effekt er så betydende, at de ligger over et valgt signifikansniveau. Vi har erfaret, at man ved denne metode i praksis ofte har flere variable end man teknisk kan håndtere i en enkelt analyse.

---

(1): Udgangspunktet er, at loglikelihood værdien udregnes for hver parameter. Den laveste af disse sammenlignes med hver enkelt af de øvrige - man udregner kvotientstørrelsen.

Uanset karakteren af det datameteriale, man vil undersøge, kommer man ikke uden om at overveje en forhåndsprioritering af sine baggrundsvariable udfra praktiske erfaringer. Under alle omstændigheder vil den endelige model være præget af de(n) person(er)s holdning til det aktuelle datamateriale.

Slutteligt - når man er nået frem til den endelige model - er man i stand til at give en prognose for en person med givne baggrundsvariable ($z^o_1, \ldots z^o_n$). Jo mere centralt den enkelte persons variable 'ligger' i forhold til baggrundsvariablene, der er observeret i materialet, modellen bygger på, desto bedre er forudsigelsen for personen.

4. OVERVEJELSER OMKRING VORES ANVENDELSE AF COX-MODELLEN.

I dette kapitel vil vi forbinde den teoretiske baggrund for Cox-modellen med dens konkrete anvendelse. Dette betyder, at kapitlet i høj grad vil handle om at kontrollere modellens forudsætninger. Endvidere præsenterer vi de overvejelser og valg, der må gøres, når der samtidigt optræder flere dødsårsager - dødsårsager, der kan være afhængige eller uafhængige af hinanden. Dette fører desuden til, at vi må forholde os til censureringsproblematikken.

a. Argumentation for at anvende en Cox-model.

I de allerede gennemgåede, mere simple, beregninger over dødelighedsmodeller focuserer man uvilkårligt på homogene grupper i populationen, hvis dødeligheder man så kan sammenligne.

Når vi i forhold til Kaplan-Meier estimater mv. arbejdede med de grafiske resultater, var det fordi de grafiske outputs var de eneste egentlige resultater fra de ikke parametriske modeller.

I vores tilfælde vil det være en urimelig opgave at sammensætte homogene grupper, da en af vores pointer netop er at undersøge forskellige baggrundsvariables indvirken på dødeligheden (det være sig til afgang uden eksamen eller eksamen).

Cox's regressionsmodel betragter ikke grupper; men enkelt individer med individuelle baggrundsvariable. En overlevelsesfunktion vil i denne ramme referere til et enkelt individs chancer for at overleve så og så længe. Når Cox's model ofte betegnes som "semiparametrisk" er det fordi den på den ene side ikke rummer nogen antagelse om overlevelsesfunktionens form (det er den ikke parametriske del af modellen), mens den på den anden side estimerer parametre i regressions-

delen, hvis estimerede værdier beskriver de enkelte variables bidrag til dødeligheden (det er den parametriske del af modellen).

Outputet af denne model er altså ikke kun de grafiske fremstillinger af S(t) mv.; men i høj grad de estimerede parametre, der indgår i den endelige model.

Cox-modellens styrke er ikke så meget at den kan udtale sig om enkeltpersoner, men at den inddrager enkeltpersoners oplysninger og udnytter disse til fulde. Dens styrke er ikke så meget de individuelle prognoser, men den kollektive erfaringsmasse, den bearbejder og leverer som parameterværdier. Alt dette vil vi vende tilbage til ved bearbejdningen af vores resultater og også senere i dette kapitel.

Nu kunne man mene, at den "bedste" model var en fuldstændig parametriseret model, men et forsøg herpå vil gøre modellen uoverskuelig og ufleksibel.
Yderligere vil en fuldt parametriseret model lægge op til en "fuldstændig" beskrivelse og forklaring" af den menneskelige virkelighed modellen anvendes på. Dette vil vi anse som problematisk. Cox-modellens fremtoning og udstråling af autoritet er for os meget passende. Den beskriver et forløb udfra de faktiske begivenheder med et ønske om at beskrive begivenhederne udfra tal-baserede personoplysninger.

I denne forbindelse står det klart, at vi ikke kan forvente at levere den endegyldige forklaring på RUC's studietider og -forløb, dels på grund af materialets beskaffenhed, men så sandelig også på grund af den meget tal-baserede angrebsvinkel, som denne type analyse nødvendigvis benytter sig af. Det, der er afgørende i den menneskelige virkelighed kan ikke udelukkende beskrives ved tal.

b. Modellens præmisser.

Vi vil her opridse de præmisser, forrige kapitel trak

frem om Cox-modellen, for at knytte nogle bånd imellem den teoretiske opbygning af modellen og den anvendelse af modellen, som vi står over for i de kommende kapitler.

De antagelser, der gør Cox-modellen så simpel, skal naturligvis stå klart for brugeren. Grundformen i modellen er

$$(1) \qquad \lambda_i(t) = \lambda_o(t) \exp(\underline{\beta} \quad \underline{z}_i)$$

Parentesen, der udregner et vektorprodukt af de to vektorer med hver i elementer, giver to forudsætninger, som brugeren skal overbevise sig om er i orden, før modellen kan opbygges på denne måde.

Proportionalitetsantagelsen. Eksponentialudtrykket er i virkeligheden et tal (eller rettere en summation over en række tal/bidrag for hver enkelt baggrundsvariabel), der bliver ganget på den underliggende dødsintensitet $\lambda_o(t)$, der svarer til den situation hvor $\underline{z} = \underline{0}$. At der således skulle gælde en så simpel relation mellem den aktuelle dødsintensitet $\lambda(t)$ og den underliggende dødsintensitet $\lambda_o(t)$, ved hjælp af de pågældende baggrundsvariable er i virkeligheden en ganske vidtgående antagelse. Kravet er, at hver enkelt baggrundsvariabels bidrag skal opfylde denne betingelse, denne proportionalitet.

I forrige kapitel så vi hvordan vi kunne bruge den integrerede intensitet som grafisk kontrol af proportionalitetsantagelsen. Der sker det, at vi trin for trin tog en variabel ud og inddelte hele befolkningen i to eller flere grupper, delt op efter den variabel, der skal testes. I stedet for én model med én underliggende dødsintensitet får vi nu to eller flere modeller, med hver sin underliggende dødsintensitet λ_1, λ_2, λ_3, osv.

Det er logaritmen til den integrerede intensitet for hver af disse grupper, der ved hjælp af computerens estimering (se formel 3.5) laver de ønskede plots.

I tilfældet med kvalitative parametre (af typen $z_i=0$ eller $z_i=1$) vil vi få den tilfredsstillelse at få tegnet kurverne, sådan at hver enkelt lodrette afstand mellem kurverne udtrykker β-værdien. Dette er imidlertid ikke tilfældet for kvantitative parametre, hvor brugeren ad hoc må indlægge nogle gruppeadskillelser: Dette skøn kan meget vel være bestemmende for, om antagelsen bliver godtaget eller ej!

Om disse grafiske kontroller gælder altså, at de er omstændige på grund af mangfoldigheden i talmaterialet, idet gruppeopdelingerne/kontrollen må foregå succesivt ved at afprøve parameter efter parameter, hver enkelt opdelt efter brugerens eget skøn. Når der er tale om kvantitative variable får brugerens valg af opdeling urimelig stor betydning.

Vi er noget utrygge ved denne metode, men kan ikke finde nogen bedre. Den udviklede numeriske kontrol (Andersen, 1982) hviler på de samme forudsætninger.

Proportionalitetskontrollen kan - om nødvendigt - betyde at man må fastholde en stratificeret (opdelt) undersøgelse, dvs. opdele modellen dér, hvor det (de) største problem(er) med manglende proportionalitet viser sig at opstå. Blot må man så overveje, om modellen nu ikke går hen og bliver for uoverskuelig til ens formål.

Men når proportionalitetsantagelsen virkelig viser sig at holde, giver modellen den fantastisk umiddelbare oplysning, at den enkelte baggrundsvariable hele tiden påvirker individet i undersøgelsen på den samme måde (med styrken $\exp(\beta \cdot z)$) i forhold til den underliggende dødsintensitet $\lambda_0(t)$.

"Alt andet lige" antagelsen. Med alt andet lige antagelsen forstås følgende: Med n parametre udtaler den første β-værdi sig om den første variables betydning for dødeligheden mellem to personer med de øvrige variable som identiske. I opbygningen af den endelige model leder

vi efter et antal variable med tilhørende parametre, der bedst muligt beskriver variationen i materialet. Som det vil fremgå af et senere kapitel, vil der for hver ny variabel ske en ændring af parameterværdierne (β'erne) for de andre variable. Det sker naturlivis, fordi at den faktiske variation ved hjælp af likelihood ligningerns maksimalisering fordeles ud på de givne parametre alt efter deres indbyrdes sammenhænge (korrelationer). Hvis f.eks. alkoholmisbrug er stærk korreleret med køn, vil udvidelsen af en model der i forvejen indeholder køn og andre variable, med alkoholindtagelse, kunne betyde en særlig stærk nedgang i parameterværdien for køn, fordi den nye parameter sluger en stor del af forklaringsværdien for køn.

Men hvis nu sygdommen ikke forholdt sig på samme måde for f.eks. køn og alder (eks: dødsrisiko størst for mænd og større med alderen); men kun forholdt sig sådan for mænd, mens kvindernes dødsrisiko mindskedes med alderen, vil modellen umiddelbart skjule denne omstændighed. Det ville jo være ganske ødelæggende, hvis det drejer sig om nogle vigtige variable.

Det er brugerens opgave at sikre, at diverse vekselvirkninger bliver estimeret og
testet mod den antagelse, at den udvidede model ikke er bedre end den gamle. Det kan man kun gøre, hvis man på forhånd (via erfaring iøvrigt) har undersøgt, hvilke vekselvirkninger, der kunne tænkes at findes i materialet.

I tilfælde med vekselvirkning vil det betyde, at de selvstændige parametre bliver udvidet med sammensatte parametre jvnf. eksemplet på side 33. Bemærk at vi ikke siger, at variablene er uafhængige selvom det skulle vise sig, at der ikke kan findes vekselvirkninger. For det vil de jo sjældent være (jvnf. en korrelationsmatrix på materialet), og det antages slet ikke.

Alt andet lige antagelsen er, trods problemerne, lige

netop den, der gør det så nemt at konkludere på de estimerede β-værdier. Hvis antagelsen er kontrolleret i et forsvarligt omfang (alle mulige vekselvirkninger kan det næppe betale sig at undersøge) vil brugeren kunne gå ind i situationen og evt. ændre på nogle af de beskrevne forhold, så at overlevelsen kan forventes at ville øges (eller evt. mindskes i forhold til RUC-studietiderne).

c. Flere afgangsårsager.

I vores materiale er det sådan, at der kan aflæses netop to afgangsårsager. Man kan faktisk kun forlade RUC med en eksamen eller uden en, sådan at den eneste censurering, der forekommer er tidscensurering, fordi mange er studerende i det øjeblik undersøgelsen stopper (pr. 1/9 1984). Vi må altså tage højde for, at der er to måder at forlade RUC på. Hvis vi blot ville undersøge, hvorlænge en studerende har opholdt sig på RUC uanset 'resultatet' af sit ophold, kunne vi blot betragte afgangene som ens (hvilket vi også kunne gøre, se senere).

Men hvis vi vil gå videre end det, må vi gøre os nogle overvejelser, idet modellens udseende nødvendigvis må tilpasses disse ændrede betingelse, eller evt. forkastes.

De konkurrerende afgangsårsager betragter vi som en achilles hæl i Cox-modellen, idet vi kunne forestille os mange konkrete undersøgelser, hvor de konkurrerende afgangsårsager vil forhindre at man med rimelighed kan anvende denne type model. Vi vil nu se på det teoretiske problem.

Man har tidligere antaget, med udgangspunkt i industrielle eksempler med serieforbundne maskiner, at hver dødsårsag havde sin egen (evt. ukendte) dødelighedsfordeling. Vi vil betragte en motors levetid - det vil sige den tid den fungerer. Motoren består af n for-

skellige dele, og hændelsen er, at den går i stå. Vi kan nu observere, hvilken fejl, der fik den til at standse. Hver af de n dele kunne være skyldige. Hver enkelt del tænkes at have sit 'holdbarhedsforløb', sådan at den slides op og på et eller andet tidspunkt, går i stykker og får maskinen til at stoppe. Det er klart, at maskinen opfører sig som en kæde: Den er ikke stærkere end dens svageste led. Hver enkelt del tænkes så tildelt en overlevelsesfunktion S_j, hvor j refererer til maskin delens nummer. Det aktuelle tidspunkt, maskinstoppet finder sted på, kan beskrives således:

$$T_i = \min(T_1, T_2, \ldots, T_j)$$

hvor T_1 til T_j refererer til de tidspunkter, hvor den j'te maskindel kunne tænkes at svigte.

Det er oplagt, at denne antagelse om underliggende dødsintensiteter kræver, at afgangsårsagerne er indbyrdes uafhængige. Det vil altså sige, at når en maskindel bliver svækket ændrer det i sig selv ikke forløbet blandt de øvrige maskindele; kun tiden (og tempoet) betyder noget.

Denne antagelse kaldes i litteraturen for antagelsen af uafhængige dødeligheder med skjulte eller latente dødeligheder for hver enkelt dødsårsag.

I det fleste lægevidenskabelige sammenhænge forekommer denne antagelse temmelig absurd. Sygdomme vil spille ind - dels direkte ved at én sygdom er en direkte følgevirkning af en anden, dels ved at een sygdom opnår at svække legemets modstandskraft så meget, at andre får lettere spil. Sygdomme er altså for det meste afhængige af hinanden.

Både antagelsen om uafhængighed og antagelse om afhængighed vil vise sig at skabe problemer for Cox-modellen. Det vil være meget rart at slippe for disse konkurrerende afgangsårsager.

Hvordan vil vi kunne afgøre, om begivenheder er afhængige eller uafhængige?

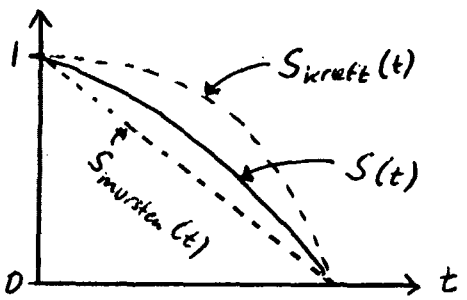
Det vil i hvert enkelt tilfælde være op til brugerne at skønne dette forhold ud fra detail kendskab til den konkrete undersøgelse. Vi kan give nogle fingerpeg.

En betingelse for at der er tale om uafhængige dødeligheder er, at sygdom j giver anledning til samme dødelighed alt imens andre sygdomme påvirker befolkningen såvel som når de andre sygdomme ikke påvirker befolkningen.

Det vil vi prøve at belyse igennem et par (konstruerede) eksempler.

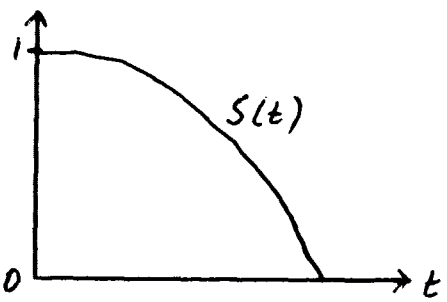
Ex. 1. Mursten og kræft. En befolkning er udsat for to døds muligheder; nemlig at dø af kræft og at dø ved at få en mursten i hovedet. I hele observationsperioden dør halvdelen af kræft og resten af at få en mursten i hovedet. For overskuelighedens skyld opdeler vi også hele befolkningens overlevelseskurve i de to gruppers:

FØR



Hvis vi nu fjerner faren for at få mursten i hovedet:

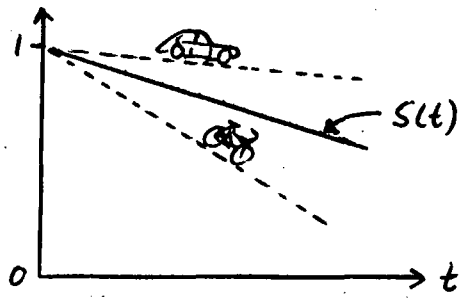
EFTER



så vil det være rimeligt at forvente at $S_{kræft}(t)$ forløber ligesom før, selvom nu alle individer dør af kræft. Eksempel 1 opfylder altså de skitserede betingelser for uafhængige dødeligheder mellem kræft og mursten i hovedet.

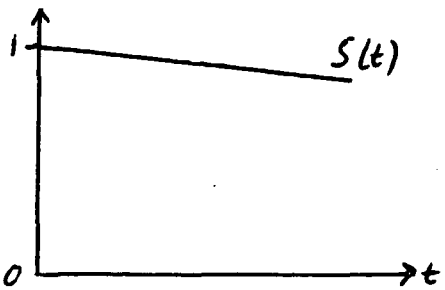
Ex 2. Bilister og cyklister. Hele befolkningen deltager i trafikken og hver enkelt individ benytter sig af bil og cykel i lige lang tid. Hvordan kan vi tænke os dødsfaldene i trafikken fordelt på henholdsvis bilister og cyklister?

FØR



Hvis vi nu fjerner muligheden for at dø i en bil (f. eks. ved at sætte hastighedsbegrænsningerne langt ned og adskille bilende og cyklende trafik) vil overlevelsen forbedres og overlevelsesfunktionen sikkert blive således:

EFTER



Eksempel 2 opfylder klart ikke betingelserne for uafhængige hændelser.

Hvad er så konsekvenserne af det ene og det andet? I stedet for én model, der kan beskrive hændelserne, må vi nu finde nogle flere; dels en estimeret overlevelsesfunktion (og tilhørende parametersæt) for hver dødsårsag og dels en sandsynlighedsfordeling for at man dør af den ene, den anden eller øvrige dødsårsager.

For først må man jo spørge: Hvad er et individs aktu-

elle chance for at dø af den bestemte sygdom, og i fald, det er den, individet vil blive ramt af: Hvordan er dødeligheden for netop dette individ fremover? Dette må gælde uanset afhængighed eller uafhængighed!

Fordelen ved denne type overlevelsesmodeller er, at alle givne oplysninger bliver inddraget i beregningerne af dødeligheden - også tal fra de individer, der er udsat for sygdommen, men som endnu ikke er døde og som kaldes censurerede personer. Om dem ved vi jo trods alt, at sidst de blev observeret var de i live. Blandt de censurerede individer kunne man også henregne de, der døde af en anden årsag, end den studerede sygdom; hvis vi er overbevist om, at den anden dødsårsag ikke konkurerer med den studerede. I så fald har denne sygdom kun interesse i censureringsøjemed. Vi har altså stiltiende akcepteret, at den syge kan rammes af andre dødsfald end det forventede.

Men ved konkurrerende dødsårsager studerer vi flere sygdomme, der hærger blandt hele observationsmaterialet. Hvis sygdommene er afhængige, har den oplysning, at patient i nu er død af sygdom a betydning for studiet af sygdom b, idet vi jo med sikkerhed ved, at patient i har overlevet sygdom b indtil da. Patient i vil således kunne betragtes som censureret i forhold til sygdom b. Men det betyder for os at se faktisk, at man påstår, at patient i ligesågodt kunne have været død af årsag b som af årsag a til det pågældende tidspunkt. Ved at inddrage alle patienter i sygdom b's risikogruppe, inddrager vi også alle patienters levetider i beregningen af overlevelsessandsynligheder mv. Brøken $\frac{t}{n}$, hvor t er den samlede levetid i befolkningen og n antallet af døde (af den studerede sygdom) øges på denne måde: Overlevelsen forbedres.

Og det vil være fuldkommen rimeligt, hvis sygdommene er så meget afhængige. Hvis man målte på lungekræft, mavekræft og lymfekræft, ville den samlede farlighed

af kræft blive voldsomt overdrevet, hvis man ikke tog hensyn til, at den ene type kræft meget let kan slå over i den anden. Regnede man ikke her de konkurrerende dødsfald med som censurerede ved beregninger for de andre kræftsygdomme ville man uden tvivl fejlvurdere farligheden af disse sygdomme i forhold til helt andre.

Teknisk set regner man altså de andre dødsfald fuldt ud med som censurerede ved afhængighedsantagelsen. Hvis man er i tvivl om afhængighed/uafhængighed er det altså ikke 'sikrere' at vælge antagelsen om afhængighed, idet man hermed ikke kan undgå at medtage antagelsen om fuld afhængighed.

Men afhængighedsantagelsen er faktisk en meget bekvem antagelse i forhold til modelbearbejdningen, her vil man kunne bruge

$$(2) \qquad \lambda_{ij}(t) = \lambda_0(t) \exp(\underline{\beta}_j \underline{z}_i)$$

hvor j referer til sygdommen og i refererer til den enkelte patient. Hver af de j beregninger inkluderer altså de andre dødsfald som censureringer og naturligvis også de egentlige censureringer.

Hvis man derimod antager at sygdommene er uafhængige, står man overfor det problem, at opdele befolkningen efter hvilken sygdom, det er, de kan dø af. Ligesom hvis vi havde en befolkning af kun hvide og sorte mennesker, der blev udsat for to forskellige sygdomme: En der kun ramte sorte folk og en der kun ramte hvide folk. Ønsket var nu at dele folk op efter hudfarven. Det er klart, at vi observerer hvem der rent faktisk dør af den ene sygdom og af den anden. Men vi ser ikke hvem, der er i risikogruppen - og ækvivalent hermed: Hvortil vi beregner censureringerne.

Censureringsmekanismen skal være tilfældig, dvs. den må ikke fungere sådan, at censurering hyppigt og metodisk indtræffer lige inden det tidspunkt en bestemt sygdom alligevel ville have krævet sit offer (eller

en anden systematisk sammenhæng).

Ex 3. Vi har en befolkning, hvoraf halvdelen er læger. Befolkningen bliver udsat for kræft og vi censurerer folk, der dør af andre (såkaldte uvedkommende) årsager. Imidlertid observerer vi en lang række selvmord blandt lægerne, som iflg. konventionen henregnes til censureringerne. Men så vil censureringsmekanismen være urimelig, fordi vi kanformode, at lægerne vælger at tage deres eget liv fremfor at dø af kræft.

Ved konkurrende afgangsårsager må censureringerne ikke foretages efter en prognose over, hvilken gruppe (sygdom) vedkommende måtte forventes at havne i (dø af) fremover. Tildelingen af censureringerne skal ikke være individuelt begrundet. Censureringerne siger netop, at person i har været udsat for sygdom j i t måneder og tabes af syne nu, men hvis person i var død, ville denne være død af årsag j.

I det øjeblik vi har formået at dele befolkningen op i "sorte" og "hvide", er der intet i vejen for at beregne de latente dødeligheder indenfor hver enkelt gruppe. Men netop denne opdeling er i de fleste lægevidenskabelige tilfælde umulig og urimelig.

I vores materiale kan vi ikke umiddelbart se nogen samenhæng mellem dét at få eksamen og dét at udgå uden eksamen. Det er ikke sådan, at man erfaringsmæssigt ligeså godt kunne få eksamen selvom man rent faktisk får afgang. Modsat behøver man nu ikke at stå specielt fjernt fra at få en eksamen. Og hvis man står foran en eksamen behøver man ikke at stå langt fra en afgang uden eksamen; men kan godt være det.

Men det vil være utroligt voveligt at påstå at der ingen afhængighed findes blandt materialets "dødsårsager"; men når valget står imellem fuld afhængighed og ren uafhængighed, tvivler vi ikke på at uafhængighedsantagelsen ligger tættest på sandheden. Konsekvensen af dette valg indebærer bl.a. en klar fordel i

de gennemsnitlige levetider: Paradoksalt nok betyder dødsfaldene i den ene gruppe, at levetiderne i den anden gruppe fremstår som længere, gennemsnitlig set, så længe vi opererer indenfor den "afhængige" models rammer. Dette paradoks findes ikke i den "uafhængige" model. I den afhæn gige model vil summen af levetiderne være s+d, hvor d er summen af censurerings-levetiderne (dødscensureringerne) fra folk døde af den konkurrerende dødsårsag. $\frac{s}{n} < \frac{s+d}{n}$.

Når dette valg nu er truffet står vi så over for det næste problem: Hvordan kan vi opdele vores befolkning? Vores censureringer består af aktive studenter, der var aktive (eller havde orlov) pr. 1/9 1984. Hvordan skal de opdeles?

En overgang overvejede vi at lave en kvantitativ størrelse, der kunne være beregnet således: Studietid/studietrin. Et lavt tal ville tyde på gode eksamens-chancer. Men sådanne metoder vil være censureringer tildelt efter individuelle prognoser, og kan ikke accepteres. For det står klart, at der ikke må fordeles til grupper efter, hvad der kunne være sket fremover med hver enkelt; prognosemetoden vil derudover være utrolig speget og vil ikke høre hjemme på dette tidspunkt af undersøgelsen.

Vi må koncentrere os om nogle stokastiske fordelingsmekanismer mellem de censurerede. Vi kan vælge at fordele efter en samlet nøgle, der hedder, at antallet af censureringer tildelt til eksamensgruppen skal svare til den andel, eksaminerne indtil nu udgør i f.t. afgange uden eksamen. En ren tilfældighedsalgoritme skulle søge for den konkrete fordeling af enkeltpersonerne.

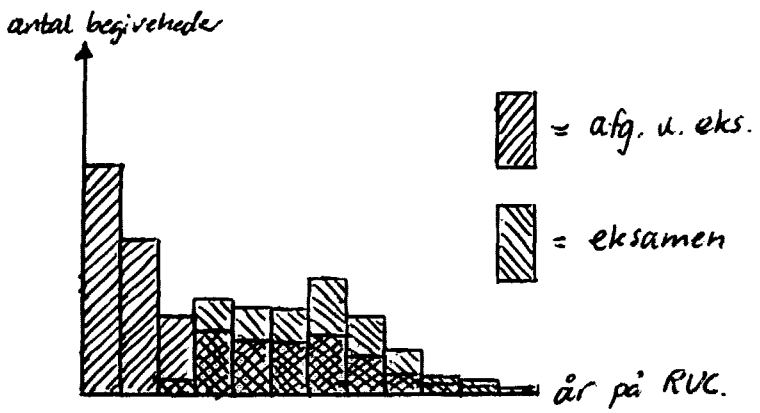
En sådan mekanisme vil dog fordele de aktive på en ganske mærkelig måde. For når man har gået 1 år på RUC er man ikke udsat lige meget for "sygdommen" eksamen og "sygdommen" afgang uden eksamen. Selvom man

selvfølgelig ikke véd, hvad man faktisk vil gå hen og "dø" af, véd man vel nok at man næppe indenfor det første "leveår" i undersøgelsen kunne være "død" af "sygdommen" eksamen. (Denne antagelse er ikke individuel.)

Vi skal altså lede efter en mekanisme, der "skævfordeler" de censurerede sådan, at i hvertfald de første 4 leveår ikke bliver særlig udsat for "sygdommen" eksamen. Hvis man nemlig tænkte sig, at vi ophobede en masse eksamens-censureringer inden den første eksamen ville den første eksaminant hive et jordskred af censureringer med sig - da begivenhederne (dér hvor overlevelseskurven bevæger sig i y-aksens retning) noteres ved egentlige dødsfald. Det er jo også derfor, at vi i et tidligere kapitel forudsatte, at censureringer - i store træk - skal være pænt tidsmæssigt fordelt blandt dødsfaldene.

Vi vil i stedet vælge at beregne de forholdstal, de faktiske eksamenspersoner og afgangspersoner (delt op efter hele år på RUC) tilvejebringer, og bruge forholdstallet på de aktive årgange. Forholdstallet illustreres via et histogram af følgende art:

Figur 4.1 Antal "begivenheder" fordelt efter studietid



Andel af censureringer til eksamensgruppen =

$$\frac{\text{faktiske eksaminer}}{\text{alle afgange}}$$

Hermed synes vi ikke, vi har sagt noget, der ligner en prognose; men vi har til fulde udnyttet de faktiske oplysninger om dødelighedsfordelingerne på RUC fordelt efter samlet studietid. Denne fordeling rummer data fra alle årgange på RUC, og eventuelle tendenser til ændringer af fordelingen hen over kalenderårene kan selvfølgelig tænkes. Men vi vil nødig gøre tildelingsmekanismen for kompliceret, så vi holder os til den foreslåede metode, og vender selvfølgelig tilbage til de konkrete tal i et senere kapitel.

d. Hvilke modeller vil vi etablere og hvorfor?

A. Vi ville have lavet S(t,**z**), der sammenlægger de to dødsårsager til én, og som derfor udtaler sig om hvor mange folk, der vil befinde sig på RUC fremover. Hvis man f.eks. lukkede fag og hovedområder, lavede særoptag på visse studier, kan modellen udtale sig om de nye behov for eksempelvis lærerkræfter, lokaler og udstyr.

B. Vi ville lave $Q_j(t,\mathbf{z})$, der skulle være af typen: Hvad er mine chancer for at jeg efter t års ophold på RUC med mine baggrundsvariable får en eksamen, er gået ud eller stadig er studerende.

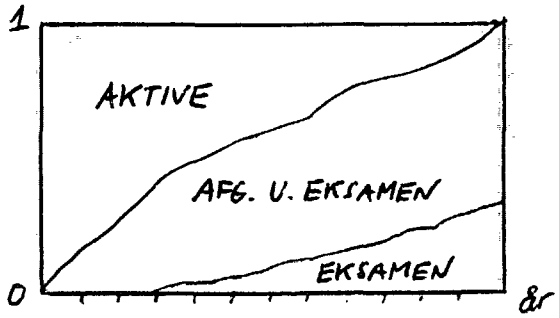
Udtrykket beskriver sandsynligheden for at dø af årsag j inden tiden t og ser således ud:

$$Q_j(t) = P(\text{død inden tid } t \text{ af årsag } j) = \int_0^t S(u)\lambda_j(u)du$$

dvs. en "summation" over de små intervaller du over chancen for at være overlevet alle sygdomme gange risikoen for til hvert enkelt tidspunkt at dø af netop sygdom j. S(u) er den samme størrelse som i A), mens $\lambda_j(u)$ er dødsintensiteten i gruppe j (hvilket skyldes, at vi jo arbejder med uafhængige dødeligheder). Udfra

et passende valg af parametre vil vi via den iøvrigt benyttede EDB-model selv kunne foretage integrationen og udregne denne model.

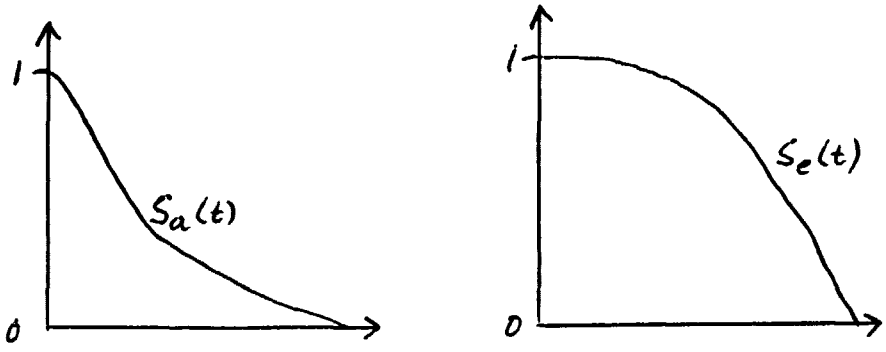
Til tiden t vil $Q_1(t) + Q_2(t) + S(t) = 1$ (med to grupper), sådan at vi, med kendskab til $Q_1(t)$ og $s(t)$ kan lave følgende diagram:



Modellen hér har særlig interesse for den enkelte studerende, der herigennem kan tilrettelægge sit (fortsatte) studium med skyldigt hensyn til de chancer, tidligere erfaringer oplyser én om.

Modellen (og især parameter-værdierne) fortæller planlæggeren og RUC-administrationen, hvor man skal sætte ind for at øge chancerne for en eksamen.

C. Vi vil lave $S_j(t,\mathbf{z})$, som er de to latente overlevelsesfunktioner.



Hvis man tror, man vil få en eksamen, benytter man sig af "eksamens"-modellen: Sætter sine prognostiske værdier ind, får et tal, der angiver ens proportionalitetsfaktor i f.t. baggrundspersonen/erne og kan via EDB få udtegnet sin overlevelsesfunktion.

Men man kan også kigge på "afgangs"-modellen og få sine værste anelser beskrevet.

Parameterværdierne fortæller noget om de faktorer, der påvirker eksamenstempoet/afgangstempoet: Altså "dødeligheden" eller "farligheden af sygdommen".

Vi får mulighed for at udregne medianer for udvalgte grupper, at tegne overlevelseskurver (evt. med sikkerhedsintervaller) og meget mere.

5. MATERIALETS BESKAFFENHED.

Dette kapitel skal belyse den situation, vi stod i: Med et dæmrende teoriapparat og en barsk og besværlig virkelighed. Hvis vi på forhånd havde vist, hvilket slæb det ville blive, var vi nok næppe begyndt på dette tal-materiale. Men nu er det gjort.

a. Hvorfor vælge matrikeloplysninger?

Planlægningen af data-indsamlingen forløb samtidig med at vi satte os ind i, hvordan overlevelsesdata kunne behandles. Det var derfor naturligt at søge efter personoplysninger fra en "hel" befolkning over et vist åremål. Et sådant materiale ville ikke kunne indfanges via andre metoder end gennem Universitets Centrets egen studenter-matrikel. Vi vidste også, at ingen andre tidligere har lavet en personbundet studieforløbsregistrering.

Vi havde forventet, at matriklen kunne levere ensartede og gerne meget informative oplysninger på hver enkelt studerende, der kunne oplyse noget om studietider, afgangsårsag, osv; men vi havde ikke på forhånd noget detailkendskab til hverken oplysningernes art eller tilgængelighed.

Vores interesse er ikke umiddelbart dét at indsamle data, og vi havde en forhåbning om, at dataindsamlingen skulle kunne forløbe relativt smertefrit, f.eks. via EDB.

Vi var allerede på et tidligt tidspunkt klar over, at den påtænkte analysemetode meget vel ville kunne rumme store anvendelsesmuligheder blandt RUC's administration. Det betyder, at materialet skal kunne korrigeres og ajourføres relativt let. Og det betyder,

at koder og oplysninger let skal kunne anvendes af andre end os selv.

b. Hvad ville vi gerne vide?

Efter en brainstorm i starten af forløbet fandt vi frem til utrolig mange forhold, der kunne påvirke et studieforløb. Mange af disse forhold ville dog i vort regi være umulige at inddrage. I flæng kan nævnes studenternes økonomiske situation, afstand fra hjem til RUC, eventuelle børn, adgangsgivende eksamen (evt. med karaktergennemsnit), forældres uddannelse, indtægt etc. Udover disse eksterne omstændigheder kunne vi også godt tænke os at vide noget om nogle mere RUC-interne forhold: Vejlederassistance, gruppestørrelser, studenterpolitisk aktivitet, evt. dumpninger til eksaminer, fagintegrerede projekter, kursusdeltagelse osv, osv.

Det må være rigtigt at opstille nogle visioner for undersøgelsen, selvom man på forhånd kunne sige sig selv, at meget lidt af det her nævnte er realistisk at indsamle data om. Her spiller netop kravet om komplette data ind - og det må betyde, at vores undersøgelse måske kan komme til at pege på nogle aspekter, som dernæst kan belyses nærmere med andre metoder; f.eks. spørgeundersøgelser og sociologiske undersøgelser.

c. Realiteterne.

Det korte og det lange blev, at vi fik adgang til alle data på samtlige RUC-studenter i perioden 1972-1984, der indeholder følgende oplysninger (variable):

Personnummer, studiestart, basis-retning, studietid

på basis, aktuelle overbyningsfag, antal studieskift på OB, orlovstid, studietrin, opgørelsestidspunkt og opgørelsesstatus.

Personnummeret giver automatisk studentens køn og sammen med studiestarten studentens alder ved studiets begyndelse.

Vi har bevidst nedtonet selve basis-forløbet (ex: skift mellem basis-uddannelser, andel af orlov hér, nøjagtig tidsregistrering), idet vi fandt, at de fleste problemer og forskelle, RUC skaber for studenterne, opstår på overbyningsuddannelserne. Frafaldet på basisuddannelserne er nok domineret af helt generelle studenter-problemer fremfor overvejende RUC-specifikke årsager. Denne vurdering af basisuddannelsens betydning kunne godt have ført til, at vi kunne have nøjedes med at se på OB-uddannelserne og startet undersøgelsen hér. Men så langt syntes vi derimod ikke vi ville gå, sålænge ingen véd mere om forholdene. Vi ville ikke gerne forudsætte mere, end rimeligt var.

d. Hvad har variablene at bidrage med?

Køn og alder er gode demografiske oplysninger og er selvfølgelig med i enhver undersøgelse af denne art. Basisretningen er vigtig for at kunne vurdere på basisuddannelsernes bredde og kvalitet, og rummer mange tolkningsmuligheder i forhold til OB-fagene. Studiestarten kan sige noget om tidsmæssige ændringer på studieforholdene. De aktuelle OB-fag er selvfølgelig mest udtryksfulde i tilfælde af en eksamen. Det (nogenlunde) frie valg af OB-fag (2 fag blandt gymnasielærerfagene) og ret udbredte studieskift betyder, at denne variable er lidt svær at greje. Hertil skal tilføjes, at vi kun har noteret antallet af skift og ikke mellem hvilke fag, de er foretaget. Vi har

valgt at nøjes med at registrere den aktuelle fagkombination, dvs. den kombination vedkommende havde på opgørelsestidspunktet. En lang studietid "belaster" således kun de(t) aktuelle fag. Men det hører også "med" i et studieforløb, at der begås fejltagelser og ændres meninger undervejs. Og når vi samtidig registrerer antallet af studieskift, synes vi, at usikkerhederne om studieaktivitet (hvornår og på hvilke fag) står i rimelig proportion til vores data.

Orlovstiden indgår naturligvis som en størrelse, der trækkes fra studietiden, så netto-studietiden lader sig regne ud; men den har derudover sikkert også betydning for studietiden.

At vi har oplysninger om studietrin er i virkeligheden ret tilfældigt. Oplysningerne findes kun i de (håndskrevne) indberetninger til Danmarks Statistik. Studietrinstilvækster bruges i budgetplanlægningen på universiteterne. Eksamenskontoret påfører efter hver eksamenstermin studenternes resultater på studenterbladene. En gang om året (siden 1975) registreres studietrin og ændringer i fag og studiestatus og indsendes til Danmarks Statistik. Kopier af disse håndførte journaler er hovedkilden til vores oplysninger.

I bilag A findes en oversigt over, hvordan studietrinnene er registreret i vores undersøgelse; men kort fortalt går de fra 0 (på basis) til 6 (lige inden kandidateksamen).

e. Vurdering.

Matrikeloplysningerne rummer (naturligvis) den afgørende svaghed, at de er officielle. Dvs. at de egentlig ikke beskriver det faktiske studieforløb: Hvornår er folk studieaktive? - kun når de er registreret som "aktive"? eller måske også når de har "orlov"? Administrationens ønske om præcis registrering står

mange gange i modstrid med de studerendes faktiske gebærden sig.

Men vi er ikke stødt ind i éntydige tendenser, der kunne berettige til konkrete advarsler og korrektioner til vores tal.

En anden omstændighed, der komplicerer billedet via matrikeloplysningerne er tilgangen af studenter udefra, der typisk kommer med ét helt færdigt fag, eller dele heraf, som godskrives udfra en faglig vurdering i de pågældende studienævn. Det betyder, at disse folk gennemlever et kortere studieforløb på RUC inden en evt. eksamen. Den vigtigste gruppe af disse personer kan findes under betegnelsen "ej basis" udfor basisretningen. Denne gruppes betydning vil blive vurderet særskilt flere steder i de følgende kapitler.

f. Bureaukratiske problemer.

Det viste sig faktisk, at vi blev nødt til at opbygge vores eget register baseret på cpr-numre. Det havde vi gerne undgået, hvis det havde været muligt at tappe de ønskede oplysninger direkte fra matriklens egen data-base (der iøvrigt befinder sig på Københavns Universitet). Det ville uden tvivl have været muligt, hvis RUC-administrationen og Registertilsynet havde villet det. I så fald krævede det et dataprogram, der kunne hive de rette ting frem. Men myndighederne i denne affære turde ikke overlade os passwords og tilladelser; men var villige til at udlevere ligeså mange udskrifter, vi måtte ønske. Så i stedet for et data-program, måtte vi igang med en særlig art arkæologi: Studiet af gamle data-udskrifter og deres indbyrdes sammenhænge og kvaliteter.

Vi beklager i princippet ikke RUC-administrationens forsigtighed eller de regler, Registertilsynet har opstillet. Det er både fornuftigt og klogt at have

god kontrol med persontilknyttede oplysninger. (Se iøvrigt brevveklingerne i bilag B).

Denne noget besværlige metode med at opbygge et register indebærer naturligvis også, at nye fejlmuligheder dukker op. I det omfang vi har stødt ind i problemer har vi kunne få matriklen til at taste vedkommende personer ind på deres terminal og levere os de ønskede oplysninger. Det konkrete samarbejde med personalet på matriklen har iøvrigt fungeret upåklageligt.

g. Dataindsamlingsproblemer.

Selve volumenet af dataene var det største problem ved denne (omstændige) metode. Metoden måtte blive, at vi så at sige rekonstruerede RUC's historie set fra matriklen, ved på den måde at få alle med i materialet. I princippet har vi fulgt alle studenter fra indskrivningen til afgangen eller opgørelsestidspunktet den 1/9 1984, og en gang om året ajourført hver enkelt students oplysninger.

Dette voldsomme bogholderi måtte naturligvis foretages via EDB, og efter vores krav og ønsker udformede vores vejleder et pascal-program, der med de nødvendige justeringer var meget velegnet til opgaven.

Dette betød, at der måtte økonomiseres voldsomt med antallet af indtastninger, og at simple fejl og misforståelser ved selve indtastningen måtte søges at minimeres. En lang række kontrolmekanismer blev indbygget og en mængde konventioner måtte praktiseres (se bilag A), men nye problemer dukkede ofte op langt henne i arbejdet.

Noget af det værste hovedbrud fik vi nok ved at betragte materialets uensartethed. Det viste sig, at vi måtte oprette registret med udgangspunkt i 3 for-

skelligt opbyggede matrikeludskrifter fra 1974 og 75, der var de eneste kilder til årgangene 1972 - 74. I 1975 startede indberetningerne til Danmarks Statistik, men skiftede desværre fagkoder i 1977, hvorefter den er blevet ført ensartet. Hvordan vi stykkede oplysningerne sammen findes i bilag A.

Dette betød, at vi måtte bruge 3 fag-koder, hvortil en oversættelsesmekanisme er indbygget i vores register-EDB-program.

Det er langtfra altid, at RUC-studenterne følger de foreskrevne studieforløb. Nogle går f.eks. ud fra RUC og vender tilbage adskillige år efter (studiestart fastholder vi og giver orlov i den mellemliggende periode), tek-sam, socionomi og forvaltning søges kombineret med hinanden og med medie på trods af studieordningen, og der er en kort og en lang forvaltningsuddannelse.

Sådan opstod der mange ting, der måtte besluttes en procedure for. Disse fejlmuligheder og mere banale fejl har vi kontrolleret og søgt rettet på forskellig vis.

Der opstod undervejs i indtastningen en lang række personer, der enten virkede mærkelige eller bare ikke blev nævnt i de senere indberetningsbøger, selvom de hos os stod som aktive studerende året inden. Vi påførte udfor undersøgelsen i 1984 alle disse personer en særlig kode, der betød, at disse personer blev skrevet ud på papir og afleveret til kontrol på matriklen.

Efter denne afslutning på selve data-indsamlingen lavede vores vejleder et program, der omformede registret til en egentlig data-base (pascal-fil) med de oplysninger udfor hver enkelt student, som vi ønskede til vores statistiske model. Hér var det muligt at udskrive alle med meget lange studietider (netto over 7 års studie), meget lange orlovtider (over 5

år), lange basisstudier (over 5 år) og negative studietider(!). Denne kontrol fik da også bragt en del mærkelige ting på bordet som blev rettet.

Endelig fik vi skrevet 25 tilfældige personer ud og fik matriklen til at kontrollere disse. Udaf de 225 fejlmuligheder fandtes kun 6 fejl, hvoraf ingen var særlig grelle. Det giver knap 3% fejl, hvilket vi synes er acceptabelt.

Slutteligt slettede vi cpr. numrene og lavede vores egen nummerering.

h. Hvad kan tallene bruges til?

En lang række simple optællingsstatistikker er blevet gjort mulige med denne database. Sådanne optællinger blev indtil 1981 gjort i hånden, hvilket da også bevirkede at denne form for statistik blev opgivet.

Pascal-sproget forekommer os meget simpelt at bruge til disse optællinger og en del mindre programmer er blevet udført til gavn for det videre arbejde med Cox-modellen og til gavn for RUC-administrationen (v. Henrik Thorsen).

Tallene er ikke specielt gode som individuelle data. Dertil er fejlprocenten alligevel for betydende og detailrigdommen for ringe, mens den for gruppers vedkommende må anses for pålidelig nok, da vi ikke har kunnet afsløre systematiske fejl. En detail-oplysning udfra vores tal er altså ikke forkert bare fordi den kun omfatter få personers bagvedliggende viden - for enhver oplysning er faktisk den mest rigtige, da det er komplette data, der ligger til grund for optællingerne. Men i prognoseøjemed er små materialer altid problematiske - og det er da også derfor, at vi senere i diskussionen om overbygningskombinationerne må sætte grænser for hvor få personer, vi kan ba-

sere en parameter på. Hvad de øvrige oplysninger angår, benytter Cox-modellen alle personers data, da alle jo har en alder, et køn, et antal orlovsmåneder, studieskift osv. Kun ved basisuddannelserne deler vi folk op (ligesom ved OB); men her er grupperne (4 stk.) meget store.

6. INDLEDENDE BEARBEJDNING AF MATERIALET.

Ved hjælp af COMPAS PASCAL programmer kunne vi nu skaffe os oplysninger, der skulle afhjælpe flg. behov:

1) et generelt overblik over de indsamlede tal,
2) en simpel af- eller bekræftelse af forskellige sammenhænge ml. studietiderne og de personlige data til brug for etableringen af en fornuftig Cox-model senere.

Det er opgaven i dette kapitel at tilrettelægge de variable, således at vi derefter kan lade et EDB-program arbejde videre med tallene.

Vi har holdt konsultationer med Henrik Thorsen for at bruge de "officielle" erfaringer på RUC.

Som vi nævnte det i et tidligere kapitel er vi nødt til at overveje mulighederne for vekselvirkninger i materialet. Men da der kan tænkes et meget sort antal af disse, er det nødvendigt at sondere terrænet inden vi går over til kostbar brug af EDB-tid.

a. Statistikkerne.

Den første statistik er en gennemgang af, hvordan de studerende, vi har i databasen fordeler sig på statusgrupper.

Figur 6.1. Opdeling af studenterne i fire statusgrupper

Antal studerende i alt på RUC	6116
Antal eksamener	1231
Antal afgang uden eksamen	2979
Antal aktive	1843
Antal orlov	63

Opgjort 1/9-84.

Det skal her bemærkes, at alle, der ikke har nået at være tilmeldt RUC i mere end 1 måned, ikke er med. Der vil altså, hvis vi havde optalt samtlige tilmeldte på RUC i tidens løb, være flere end 6116. Statistikken viser, at af samtlige, der er gået ud fra RUC, har $\frac{1231 \cdot 100}{(1231+2979}$ % = 29 % fået eksamen - resten er gået ud uden eksamen (afgang).

De følgende statistikker, vi vil vise, udspringer dels af den "almindelige" formodning om studiestrukturen på RUC og dels af nogle forhold vi uden nærmere begrundelse finder interessante. Disse statistikker kan så give os inspiration til at inddrage bestemte variable i Cox-modellen. Eksempelvis har det altid været et ubekræftet rygte, nærmest en kendsgerning blandt RUC'ere, at basisterne bliver yngre og yngre, når de starter.

Figur 6.2.Gennemsnitsalder ved studiestart for hhv. basister(1) og samtlige studerende(2).

I denne statistik optræder der to værdier for alderen, en for samtlig studerende, og en for basister. Denne adskillelse vil ofte optræde i resten af de følgende statistikker, og den skyldes, at der, som det kan ses, er forskel på de to grupper. Forskellen skyldes, at dem, der kommer ind på RUC uden basis på en eller anden måde har gjort sig fortjent til at undvære basis. F.eks. kan det være folk, som har studeret på andre universiteter, DTH, DIA eller lignende, hvilket betyder at de oftest er ældre end basisterne.

Ikke alene vil de fleste "udefra" springe basis over, men vil ofte nøjes med ét fag eller måske dele af et.

Udfra kurverne kan man ikke umiddelbart få den gængse formodning om lavere gennemsnitsalder ved studiestart bekræftet - selv om tendensen efter 1979 går klart mod lavere startalder.

Figur 6.3. Gennemsnitlig studietid indtil eksamen på samtlige OB-fag.

Mænd:	606	med	66,9	mdr.	studietid
heraf	519	-	72,3	mdr.	- incl. basis
	87	-	34,1	mdr.	- uden basis
Kvinder:	625	-	52,3	mdr.	-
heraf	538	-	56,2	mdr.	- incl. basis
	87	-	28,0	mdr.	- uden basis

Ligesom før, hvor vi kunne ræsonnere os frem til, at startalderen for ej basister er større end for basister, kan vi i denne gruppe forvente at den gennemsnitlige studietid er kortere - mindst 24 måneder (basistiden).

På statistikken kan det også ses, at studietiden for kvinder er kortere end for mænd. Dette kan man forvente, idet langt flere kvinder end mænd læser socionomi, som er et kort studie - normeret til 3½ år.

Et interessant spørgsmål vil være om der er forskel på kvinders og mænds studietid på de forskellige overbygningsuddannelser. Vi har her for overskuelighedens skyld kun lavet fire grupper, idet vi ganske naturligt har slået alle gymnasielærerfagene incl. medie og psykologi sammen i en gruppe, da alle disse fag på en eller anden måde skal læses to og to. De resterende tre grupper er socionomi, teknologisk samfundsvidenskabelig planlæggeruddannelse (tek-sam) og forvaltning, der læses enkeltvis.

Figur 6.4. Antal eksamener og gennemsnitsstudietid på gymnasiefag, socionomi, tek-sam og forvaltning.

	Basister		Ej basister	
Gymnasiefag				
Mænd	202	87,4 mdr	36	41,0 mdr
Kvinder	97	84,3 -	27	33,5 -
Socionomi				
Mænd	152	49,1 -	31	21,5 -
Kvinder	394	47,0 -	45	21,4 -
Tek-sam				
Mænd	103	73,9 -	10	48,7 -
Kvinder	32	75,7 -	6	38,7 -
Forvaltning				
Mænd	58	78,0 -	10	34,4 -
Kvinder	10	81,6 -	9	37,8 -

Derudover er der 5 kvinder og 4 mænd, der har eksamen i en kombination mellem et gymnasiefag og et ikke-gymnasiefag.

I ingen af de fire grupper er der markante udsving i studietiden for hhv. kvinder og mænd. Desuden ses, at forskellen mellem basister og ej basister lige som før er meget tydelig.
Tilmed kan det konstateres, at kvinderne på nogle uddannelser er hurtigere end mændene, mens det omvendte

er tilfældet andre steder. Det kan meget vel vise sig, at variablen køn ikke i sig selv er signifikant p.gr.a. de modsatrettede tendenser i fagene taget under ét; men udfra denne tabel har vi et argument for at prøve en vekselvirkning mellem køn/(tek-sam, forvaltning) og køn/resten af fagene.

Nu har forskellen i studietid mellem basister og ej basister været tydelig, hvilket kunne give et fingerpeg om, at det er en fornuftig variabel. Spørgsmålet er nu om de tre basisretninger også er fornuftige variable.

Figur 6.5. Studietiden indtil eksamen afhængig af basisvalg.

Nat	124	79,8 mdr.
Hum	313	63,3 -
Sam	620	61,5 -
Ej bas	174	31,1 -

På ovenstående figur kan man se, at SAM- og HUM-basister har nogenlunde lige lange studietider. For at få et bedre grundlag for at bedømme om denne opdeling er rimelig har vi lavet tre statistikker over henholdsvis geografi, historie og biologi med vilkårlige andet fag. Disse tre fag repræsenterer hvert af de tre hovedområder.

Figur 6.6. Studietiden indtil eksamen på geografi, historie og biologi + 2.fag afhængig af basisvalg.

Nat	27	88,1 mdr.	
Sam	47	86,9 -	
Hum	9	84,8 -	Geografi
Ej bas	18	39,9 -	
Nat	0	0 -	
Sam	72	87,8 -	
Hum	50	90,5 -	Historie
Ej bas	14	48,4 -	

Nat	34	85,3	mdr.
Sam	0	0	-
Hum	2	99,0	-
Ej bas	5	44,6	-

Biologi

Disse tre fag er valgt ud, fordi man kunne tro, at basis-uddannelsens virkning på OB-fagene hér tydeligst ville kunne ses.

Et gennemgående træk er, at studietiderne indtil eksamen faktisk er ret ens. Det bedste billede ses for geografi's vedkommende. Hér er der nemlig en temmelig pæn repræsentation fra alle 3 basisuddannelser.

Historie og biologi udmærker sig ved at være mindre brede, end man kunne have troet, idet kun 2 basisuddannelser har leveret studenter, der siden har bestået en eksamen. De to med HUM-BAS på biologi har godt nok brugt væsentligt længere tid end de 34 fra NAT-BAS; men 2 udaf i alt 41 er lige lidt nok materiale at basere nogen vekselvirkningsantagelser på.

Skulle vi have undersøgt vekselvirkninger mellem basis og OB-hovedområder ville det have været muligt ved at kombinere hver enkelt basis (incl. ej bas) med hver af de 3 hovedområder, dvs. 12 nye variable. Men på baggrund af oversigten fra geografi, historie og biologi finder vi ingen grund til denne ulejlighed.

Man kunne tro, at aldersforskelle ikke betyder det samme alle steder på tidsskalaen; dvs. at springet fra 18-22 år betyder mere end springet fra 30-34 år for studietiden. Derfor vil det være rimeligt at prøve at indføre log(alder)de steder, hvor alderen optræder.

En anden antagelse, vi har fundet det interessant at undersøge, er, om det er afhængigt af valget af basis-retning, hvordan alderen påvirker studietiderne. Det foregår simpelthen ved at erstatte 0-1 variablen udfor basisretningerne med en 0-alder variabel udfor basisretningerne. De nye variable udtaler sig ikke

bare om basisretning, men også om alderens betydning på de enkelte basisretninger.

På sammen måde vil man kunne kombinere de forskellige OB-kombinationer med alderen.

Figur 6.7. Planlagte forsøg med de indgåede variable.

Log alder
Køn og (tek-sam + forvaltning)/Køn og (resten)
Basisretning/alder
OB-kombinationer/alder

Selv om vi hidtil har vægtet begivenheder inden en eksamen (og agter at generalisere de heraf udledte antagelser til også at gælde for de udgåede), har vi også arbejdet med de udgåede alene. Den følgende tabel er en slags "hitliste" over de mest upopulære OB-fag på RUC.

Figur 6.8. Studietid indtil afgang og frafald for ægte RUC'ere, fordelt på fag.

	Genst st. tid indtil afgang.	Antal udgåede studerende.	Studerende i alt.	Frafald i %.
Fransk	51	13	37	35,1
Engelsk	48	32	95	33,7
Historie	54	148	440	33,6
Samfundsfag	54	159	495	32,1
Dansk	46	84	288	29,2
Forvaltning	48	64	282	22,7
Geografi	50	75	352	21,3
Fysik	48	9	44	20,5
Matematik	58	12	64	18,8
Tysk	55	9	50	18,0
Biologi	45	26	156	16,7
Datalogi	51	17	112	15,2
Tek-sam	44	54	401	13,5
Kemi	49	6	47	12,8
Medie	45	24	207	11,6
Socionomi	39	76	686	11,1
Psykologi	74	3	33	9,1

Forklaring til figur 6.8.(forrige side): Disse tal vil ikke være lette at aflæse fra de planlagte modeller så derfor er de fundet frem på denne måde. Kolonnen "i alt" omfatter antallet af samtlige studerende (uanset afgangsårsag) igennem tiderne.

b. Gruppering af OB-fagene.

Et gennemgående problem ved model-tilrettelæggelsen er gymnasielærerfagenes kombinationer (gruppering). Som dataerne er nu, har hver person nogle fagkombinationer. F.eks. matematik/fysik eller matematik/historie. Hvis nu kombinationen matematik/fysik har brugt 8 år og matematik/historie 6 år på at få eksamen, i hvilket fag ligger så begrundelsen?

Lad os først forudsætte, at alle de andre variable er ens,for så kan forskellen kun skyldes kombinationen. "Skylden" kan ikke med rimelighed tillægges matematik alene, men må ses i kombination med et andet fag; det vil sige, at det er fornuftigt at betragte en persons to gymnasielærerfag som én variabel, som indeholder kombinationen.

Spørgsmålet er herefter, om vi så skal lave én parameter for hver praktiseret kombination. Eller om det forekommer for vildt at lave en generalisérbar betaværdi på baggrund af én eller få personers studieforløb.

Vi valgte at lade 5 personer udgøre den mindste gruppe inden for henholdsvis eksamensbefolkningen og inden for afgangsbefolkningen.

Figur 6.9. Eksamenspersoners gymnasie-OB-fag.

	Fag	i kombination med	eller med fag fra hv.omr. SAM	NAT	HUM
Samfv.	Geografi	medie 9 historie 17 samfundsfag 46 biologi 17 intet fag 5		1	
	Samfundsfag	historie 77 medie 11 dansk 18			3
Naturv.	Matematik	biologi 5 fysik 7	2		1
	Fysik		2		
	Datalogi		1	2	
	Biologi	kemi 8		6	4
	Kemi			2	
Humaniora	Dansk	fransk 5 medie 9 psykologi 6 engelsk 9 historie 26	1	1	2
	Engelsk	medie 5	3		2
	Fransk		2		
	Tysk	historie 7			
	Historie				7
	Psykologi	intet fag 14	3		2
	Medie	intet fag 13			

I tabellen ser vi de eksamenskombinationer, der findes, dog sådan at søjle 1 navngiver de kombinationer, der kan mobilisere mindst 5 personer. De øvrige muligheder er noteret i "opsamlingsgrupper" i søjle 2-4. Tallet 5, der bruges som grænse, er fastlagt efter et skøn over, hvad der var rimeligt at danne parametre efter.

Opsamlingsgrupperne bliver endeligt også slået sammen på den lodrette led. Disse grupper er med for at sikre, at alle personer bliver placeret entydigt i grupper - hvilket må være nødvendigt for at vægte OB-fagenes samlede effekt på studietiderne.

Dertil kommer selvfølgelig de fag, der læses alene. Medie er dog blevet gennemført med socionomi; mens andre læste kombinationer er slået sammen i det ene enkelt-fag.

En tilsvarende tabel er lavet for udgået-personernes fag-kombinationer.

For overblikkets og EDB-tidens skyld har vi fundet det rimeligt at anvende de samme OB-fag-grupper i begge undersøgelser. Det er gjort ved at udvælge de grupper, hvor mindst 5 personer figurerer begge steder. Følgende kombinationstyper er herefter udvalgt:

Bio/geo, bio/kemi, dansk/eng, dansk/hist, dansk/samf, mat/fys, geo/hist, geo/samf, hist/samf, hist/tysk, geo/alene, psyk/alene, samf/alene, medie/alene, teksam, socionomi, soc/medie, driftsøkonomi, forvaltning, et NAT-fag, et HUM-fag, 2 NAT-fag, 2 HUM-fag, SAM/NAT, SAM/HUM, NAT/HUM. I alt er der 26 fagkombinationer.

c. Tildeling af censureringer på eksamnesgruppen og afgangsgruppen.

I kapitel 4 blev censureringerne omtalt som et problem. Normalt regner man blot censureringer ind i en model som nogle personer, der udgår af undersøgelsen uden at den eftersøgte hændelse er fundet sted. Men i vores tilfælde er der to modeller; én for henholdsvis eksamen og afgang. Hvis vi bare lægger censureringen ind begge steder bruger vi oplysningerne to gange, hvilket forlænger overlevelsen mere end det burde. I det tilfælde, at vores to dødsårsager var afhængige, ville der ikke være noget problem, men de er jo som tidligere beskre-

vet uafhængige (en udførlig argumentation findes i kap. 4.c.).

Vores bud på at løse problemet er at vurdere hver enkelt censurering udfra et simpelt kriterium, hvorefter vi enten tilskriver eksamen eller afgang censureringen.

Vores kriterium er, at vi fordeler censureringerne efter hvilket afgangsårsag personen forventes at dø af. Da der endnu en gang er stor forskel på basister og ej basister deler vi op i disse to grupper. Grundlaget for fordelingen af censureringerne laves udfra forholdet mellem eksamen og afgang inden for hvert enkelt studieår i databasen. For nemmere at forstå princippet i udregningen viser vi her de første beregninger.

Blandt studerende med basisuddannelse, der har læst mellem 4 og 5 år er der NN, der får eksamen og WW, der får afgang, dvs. $\frac{NN \times 100}{NN+WW}$ % chance for at få eksamen. I figur 6.10 kan man se, hvordan tildelingen til eksamensgruppen fordeler sig mellem de aktuelle studerende pr. 1/9 1984 alt efter deres netto-studietid indtil da. (Se figur 6.10 på næste side).

Argumenter for vores kriterium er, at de studerende skal censureres efter den afgangsårsag, de har været udsat for. Husk på, at det eneste vi ved om de studerende, der er tidscensureret, er, at de ikke har fået hverken eksamen eller afgang til det tidspunkt, de er censurerede, de har altså "overlevet" de to afgangsårsager, som hver især har floreret med den kraft som procenttallene er udtryk for. Vi kunne så gå længere ned i detaljerne end bare basister/ej basister, f.eks. fagkombination; men som før nævnt mener vi, at fordelingen ikke bliver så tilfældig, som censureringer bør være.

Figur 6.10. Fordelingsnøgle til tildeling af censurerede til eksamens-databasen.

Studietid	Med basis-uddannelse: Antal studerende	Med basis-uddannelse: %-vis tildeling til eksamensgr.	Med basis-uddannelse: Tildelt antal efter tilfældighedsmekanisme	Uden basis-uddannelse: Antal studerende	Uden basis-uddannelse: %-vis tildeling til eksamensgr.	Uden basis-uddannelse: Tildelt antal efter tilfældighedsmekanisme
0-12 mdr	342	0,0	0	38	3,5	2
13-24 mdr	379	0,0	0	31	74,5	21
25-36 mdr	340	6,5	16	36	70,6	27
37-48 mdr	252	68,7	168	21	83,3	18
49-60 mdr	179	73,3	133	10	82,4	8
61-72 mdr	134	79,9	105	2		1
73-84 mdr	47	86,3	43	0		0
85-96 mdr	27	88,5	23	1		1
97-108 mdr	25	81,0	24	0	66,7	0
109-120 mdr	26	70,0	19	1		0
121-132 mdr	10	71,4	8	0		0
133-144 mdr	5	0,0	0	0		0
	1766	544 stk.	539	140	78 stk.	78

Et EDB-program i COMPAS PASCAL fordelte de 6116 personer ud i to nye databaser til brug for det videre arbejde. Tilfældighedsmekanismen ramte helt pænt det tal, der var bestemt på forhånd. Eksamensgruppen havde 1231 eksamener og fik tildelt 617 censureringer, mens afgangsgruppen havde 2979 udgåede og fik tildelt 1289 censureringer.

7. MODELLENS ENDELIGE KONSTRUKTION.

I det følgende kapitel vil vi præsentere vort EDB-program BMDP samt ret grundigt berette om vore bestræbelser på at få en endelig model tilpasset via vore data. Endelig vil vi prøve at forklare nogle af de opnåede resultater - forklaringer, der ikke er endegyldige, men kan danne udgangspunkt for en debat. I kapitlet koncentrerer vi os hovedsageligt om data fremkommet udfra folk, der har bestået eksamen - sidst i kapitlet er der dog enkelte bemærkninger om en model baseret på folk med afgang uden eksamen.

a. Valg af EDB-program.

Der findes flere brugbare programmer, der kan bearbejde et talmateriale til Cox-modellen. Caspersen et al 84 benytter det program i fortran, Kalbfleich & Prentice bragte som bilag i deres bog (1983).

Institut II har statistik-programpakken GLIM, der også har Cox-modellen indbygget som facilitet.

Et fælles problem for disse muligheder var databasens størrelse (6116 personer med knap 40 variable hver). Vi forsøgte at få overført GLIM til en anden maskine; IBM PC'er, der har større regnekapacitet. Da dette mislykkedes (p.gr.a. oversættelsesproblemer i fortran), stod vi med valget: Lave et program selv eller benytte os af RECKU (Regionalt Edb Center for forskning og uddannelse ved Københavns Universitet).

Hér stiftede vi bekendtskab med endnu en programpakke, nemlig BMDP.

Dét, der holdt os fra at bruge RECKU fra starten af,

var en lyst til selv at bevare overblikket over procedurerne - et overblik, som erfaringsmæssigt let forsvinder under brugen af store færdigpakkede EDB-programmer. Vi frygtede, at vi blev underlagt maskinens vilje, fremfor selv at styre processerne. Den frygt viste sig at være reel nok - vi kom til at bruge lang tid med overhovedet at få greb om, hvad det udvalgte program foretog sig.

Efter mere end en uges arbejde med pilot-projekter bestående af de 200 første personer (uden OB-fag) formåede vi dog at kunne styre programmet dérhen, vi ville.

BMDP rummer en afdeling (P1L), der svarer til de ikke-parametriske overlevelsesmodeller, vi har beskrevet i kap. 2; med Kaplan-Meier estimater, Nelson-estimater mv. En anden afdeling (P2L) rummer så selve Cox-modellen, der udover estimation af beta-værdier kan lave trinvis opbygning/elimination af parametre, lave tests, tegne overlevelseskurver osv. Programmet tillader også tids-afhængige variable.

Det viste sig hurtigt, at økonomien var et problem. For at spare penge (og tid) har vi måtte koncentrere os om én model (eksamensmodellen: $S_e(t)$, der er opbygget på grundlag af folk, der <u>har</u> fået eksamen). Herigennem har vi fået lejlighed til at gøre tingene tilstrækkelig omhyggeligt, - for senere at kunne generalisere metoden og erfaringerne hérfra til den model $S_a(t)$, der bygger på folk med afgange fra RUC. De øvrige modeller, vi ønskede os (nævnt i kap. 4) må udføres en anden god gang - nu er forarbejdet i hvert fald gjort.

<u>b. Overflødige oplysninger i databasen for eksamensmodellen.</u>

Da vi i det forrige kapitel tilpassede en lang række 0-1 variable til at beskrive OB-kombinationerne, be-

tingede vi os, at alle personer skulle kunne placeres i én OB-gruppe. Dog gjorde vi det sådan, at folk uden OB-fag (dvs. studerende, der endnu går på basisuddannelserne) ikke blev placeret i en gruppe; men alligevel kan findes entydigt blandt grupperne ved, at der i så fald vil være 0'er udfor alle OB-fag-grupper. BMDP ville logisk nok ikke acceptere indbyrdes afhængighed i variablene; da 1 gruppe OB-fag nødvendigvis er bestemt ved de andre. Det er ligesom, hvis vi havde én 0-1 variabel for 'mænd' og én for 'kvinder'. Den ene er selvfølgelig overflødig. Det samme gælder basis-uddannelserne. Hér er der også én overflødig parameter - og man kunne fristes til at fjerne EJBASIS. For den fjernede gruppe gælder, at der ikke vil blive estimeret nogen beta-værdi, endsige beregnes særskilt log likelihood værdi. Den afhængige variabel beskrives i stedet gennem den bagvedliggende dødsintensitet. EJBASIS skønnede vi var for vigtig en variabel til at den skulle fjernes. I stedet valgte vi at fjerne humbasis, som vi på forhånd ikke tillagde ret mange sprælske egenskaber.

Hvis vi har bare én overflødig oplysning, betyder det, at variablene er lineært afhængige. Det opdager EDB-programmet, når det skal invertere den matrice, der rummer de 2'den afledede af log likelihood ligningerne. BMDP klager i så fald netop over, at informationsmatricen er singulær, dvs. ikke kan inverteres.

Vores opdeling af data-materialet i 2 skarpt adskilte grupper - eksamensgruppen og afgangsgruppen - og dét, at alle personer placeres i OB-faggrupper betyder imidlertid, at nogle af de etablerede OB-faggrupper alligevel må overvejes særskilt i hver model. Der skete nemlig det, at eksamens-befolkningen fik tildelt censureringer i grupper, hvor ingen endnu havde taget eksamen. Det blev vi gjort opmærksom på, da BMDP ikke kunne udregne beta-værdier for grupperne samfundsfag-alene, driftsøkonomi og medie-alene. Som konsekvens heraf

fjernede vi disse tre OB-faggrupper - hvilket betyder, at de pågældende personer stadig benyttes i undersøgelsen; blot kan vi af gode grunde ikke udtale os om netop dé 3 variables betydning for eksamen.

Vi har udarbejdet et bilag om den geometriske fortolkning af regressionen og udvælgelsesprocedurerne hér omkring, samt af Newton-Raphson iterationen. Selvom vi beskriver frengangsmåderne i et mere algebraisk sprogbrug hér i kapitlerne, vil en beskrivelse i geometriske abstraktioner ofte være en hjælp - denne hjælp findes altså i bilag C. Den geometriske fortolkning er iøvrigt fuldt ud så "korrekt" som den algebraiske.

c. Detailplanlægning og præsentation af BMDP.

Følgende skema lagde vi for EDB-arbejdet:

1) etablere en model hvor alle parametre er inde, estimere beta-værdier og en værdi for log likelihood, oplysninger, der kan bruges som udgangspunkt for det øvrige arbejde.
2) undersøge muligheden for vekselvirkninger.
3) undersøge proportionalitet.
4) evt. omformulere modellen.
5) opbygning af betydelige parametre.
6) udtegning af karakteristiske overlevelsesfunktioner.

Vi vil kort præsentere vores kommunikation med BMDP, sådan som den formede sig under pkt. 1) i skemaet:

(se næste side).

Figur 7.1 Vores kommando til en BMDP-kørsel.

```
/PROBLEM TITLE IS 'STUDIETIDER TIL EKSAMEN PAA RUC'.
/INPUT VARIABLES ARE 37.
      FORMAT IS '(F1.0,F2.0,F1.0,2F2.0,3F1.0,F2.0,F1.0,F3.0,F2.0,25F1.0)'
/VARIABLE NAMES ARE NYSTATUS, 'ST.TID', KON, STARTAAR, STARTALD, EJBASIS,
                    SAMBASIS,NATBASIS,
                    BASISTID, 'ST.SKIFT', ORLOVTID, BIOGEO,BIOKEMI,DANENG,
                    DANHIST, DANSAM,MATFYS,GEOHIST,GEOSAM,HISTSAM,
                    GEOGRAFI,PSYKOLOGI,HISTTYSK,SAMF,MEDIE,TEKSAM,
                    SOCIONOMI,SOCMEDIE,DRIFT,FORVALTN,NAT,HUM,NATNAT,
                    HUMHUM,SAMNAT,SAMHUM,NATHUM.
      USE = 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,
            21,22,23,26,27,28,30,31,32,33,34,35,36,37.
/FORM TIME='ST.TID'.  STATUS=NYSTATUS.  RESPONSE = 2.
/REGRESSION COVARIATES ARE 3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,
            19,20,21,22,23,26,27,28,30,31,32,33,34,35,36,37.
      CONV = 0.0001.
      INIT = .1628,0.0,0.0,2.1781,0.1599,0.0,-.3731,-.0613,
            -.0266,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,2.1835,
            1.5410,0.0,.5628,2.9419,1.9117,0.0,1.4787,0.0,0.0,
            0.0,0.0,0.0,0.0.
/END.
```

I pkt. 1 fortæller vi hvormange variable, der indgår i filen og hvordan filen er inddelt for hver person (f.eks. betyder 25F1.0 at der er 25 oplysninger af ét tals længde og med nul decimaler).

I pkt. 2 fortæller vi navnene på variablene og oplyser, hvem af dem, vi ønsker at anvende i denne kørsel (bemærk at nr. 24, 25 og 29 er udeladt, som før nævnt).

I pkt. 3 fortæller vi, at st.tid er responstiden, at begivenheden, vi registrerer til responstiden (status) hedder nystatus og at et dødsfald (eksamen) er registreret med et 2-tal i nystatus.

I pkt. 4 defineres regressionen; de ønskede variable (selvfølgelig uden 1 og 2) og i dette tilfælde ønsker vi en specificeret convergens-værdi (relativ præcision i Newton-Raphson-iterationen) og oplyser nogle bestemte startværdier, som ellers sættes lig 0 (vi har tallene fra en pilot-undersøgelse).

d. Den fulde model.

Udskriften efter endt computerbearbejdning starter med 4 sider med oplysninger om filen, hvordan tallene ser ud med de rigtige navne osv. Herefter modtager vi en oversigt med nogle simple udregninger:

Figur 7.2 Beskrivende statistik over de indgående variable.

VARIABLE NO.	N A M E	MINIMUM	MAXIMUM	MEAN	STANDARD DEVIATION
3	KØN	.0000	1.0000	.4632	.4988
4	STARTAAR	72.0000	83.0000	76.2489	2.9201
5	STARTALD	17.0000	54.0000	24.1369	5.4801
6	EJBASIS	.0000	1.0000	.1364	.3433
7	SAMBASIS	.0000	1.0000	.4665	.4990
8	NATBASIS	.0000	1.0000	.1481	.3553
9	BASISTID	.0000	10.0000	1.7863	.8776
10	ST.SKIFT	.0000	3.0000	.1667	.4180
11	ORLOVTID	.0000	86.0000	2.9313	8.5982
12	BIOGEO	.0000	1.0000	.0141	.1178
13	BIOKEMI	.0000	1.0000	.0114	.1069
14	DANENG	.0000	1.0000	.0070	.0836
15	DANHIST	.0000	1.0000	.0222	.1473
16	DANSAM	.0000	1.0000	.0135	.1156
17	MATFYS	.0000	1.0000	.0054	.0734
18	GEOHIST	.0000	1.0000	.0200	.1401
19	GEOSAM	.0000	1.0000	.0400	.1961
20	HISTSAM	.0000	1.0000	.0611	.2397
21	GEOGRAFI	.0000	1.0000	.0049	.0696
22	PSYKOLOG	.0000	1.0000	.0097	.0982
23	HISTTYSK	.0000	1.0000	.0070	.0836
26	TEKSAM	.0000	1.0000	.1499	.3571
27	SOCIONOM	.0000	1.0000	.3425	.4747
28	SOCMEDIE	.0000	1.0000	.0043	.0657
30	FORVALTN	.0000	1.0000	.0985	.2980
31	NAT	.0000	1.0000	.0965	.0803
32	HUM	.0000	1.0000	.0043	.0657
33	NATNAT	.0000	1.0000	.0206	.1420
34	HUMHUM	.0000	1.0000	.0568	.2316
35	SAMNAT	.0000	1.0000	.0130	.1132
36	SAMHUM	.0000	1.0000	.0498	.2176
37	NATHUM	.0000	1.0000	.0314	.1744

STATJS CODE FREQUENCIES

	TOTAL	DEAD	CENSORED	PERCENT CENSORED
	1848	1231	617	.3339
MISSING	1			

Ved variablen køn udregner programmet et gennemsnit til 0,4632. Da vi tillægger variablen værdien 0, hvis studenten er en mand og 1, hvis det er en kvinde, kan man ud af gennemsnittet konstatere, at der er flere mænd end kvinder i denne undersøgelse. Endvidere ses, at variabel 28, SOCMEDIE, har et gennemsnit på 0.0043. Multipliceres dette tal med det samlede antal studenter (1848) fås antallet af studenter på SOCMEDIE: 0,0043 * 1848 = 8 studenter.

Vi får at vide, at censureringsfrekvensen er 0,3339.

Figur 7.3 Koefficienterne i den første fulde model.

LOG LIKELIHOOD = -7227.1468
GLOBAL CHI-SQUARE = 2115.37 D.F. = 32 P-VALUE = .0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
3	KØN	.1791	.0636	2.8156	1.1962
4	STARTAAR	.0014	.0142	.1007	1.0014
5	STARTALD	-.0075	.0060	-1.2484	.9926
6	EJBASIS	2.3056	.1783	12.9287	10.0305
7	SAMBASIS	.2130	.0767	2.7757	1.2374
8	MATBASIS	.4226	.1469	2.8763	1.5259
9	BASISTID	-.3790	.0595	-6.3674	.6845
10	ST.SKIFT	-.6468	.0822	-7.8701	.5237
11	ORLOVTID	-.0266	.0048	-5.5511	.9738
12	BIOGEO	-2.6462	.5782	-4.5763	.0709
13	BIOKEMI	-2.7770	.6263	-4.4341	.0622
14	DANENG	-2.4405	.5123	-3.2860	.0871
15	DANHIST	-2.8773	.5467	-5.2626	.0563
16	DANSAM	-2.2797	.5624	-4.0534	.1023
17	MATFYS	-3.1623	.6484	-4.8768	.0423
18	GEOHIST	-2.8512	.5631	-5.0638	.0578
19	GEOSAM	-2.3459	.5302	-4.4247	.0958
20	HISTSAM	-2.6603	.5234	-5.0824	.0699
21	GEOGRAFI	-.3328	.6724	-.4950	.7169
22	PSYKOLOG	-.9365	.5797	-1.6154	.3922
23	HISTTYSK	-2.3615	.6329	-3.7630	.0944
24	TEKSAM	-2.1144	.5154	-4.1027	.1207
25	SOCIONOM	.4290	.5067	.8467	1.5356
26	SOCMEDIE	-.5827	.6338	-.9195	.5584
30	FOKVALTN	-2.3676	.5186	-4.5655	.0937
31	NAT	-1.0697	.7656	-1.3972	.3431
32	HUM	-1.9972	.7684	-2.5992	.1357
33	NATNAT	-3.1318	.6103	-5.1317	.0436
34	HUMHUM	-2.6543	.5300	-5.0078	.0704
35	SAMNAT	-3.2239	.6586	-4.9028	.0399
36	SAMHUM	-2.6723	.5377	-4.9702	.0691
37	MATHUM	-3.0622	.5951	-5.1457	.0466

På ovenstående figur får vi det bedste estimat af beta'erne, for både de kvalitative- og kvantitative variable. Umiddelbart kan beta-værdierne udfor de kvalitative variable sammenlignes indbyrdes, men skal en kvalitativ variabel - f.eks. køn - sammenlignes med en kvantitaitv - f.eks. startår - må man multiplicere β værdien for startår (0,0014) med f.eks. 72.

Desuden estimeres standardafvigelserne og udfra disse, samt beta'erne fås nogle simple udregninger; dels koefficienten/S.E., der (numerisk) siger noget om den pågældende variabels 'forklaringskraft'(1) og dels exp(β), der angiver den pågældende variabels propor-

tionalitetsfaktor til den underliggende dødsintensitet. Værdier over 1 af exp(β) forøger dødsintensiteten; mens værdier under 1 formindsker den.

Ved denne (den bedste) estimering af beta'erne giver Cox's lig likelihood funktion værdien -7227,1468.

Den generaliserede scoretest - eller global chi-i-anden test - er regnet ud, og tester udfra første og anden afledede af likelihood ligningerne om $\beta = 0$. (2)

Sålænge hele befolkningen (dvs. alle 1848) regnes med, vil vi på log likelihood værdien kunne teste øvrige kombinationer. En bedre model har en højere log likelihood værdi. Ved at udvide antallet af parametre vil vi altid få en model, der beskriver virkeligheden bedre. Problemet er at afgøre om denne forbedring er signifikant - dvs. om den nye model med flere parametre forklarer virkeligheden signifikant bedre end den gamle. Tilsvarende vil man ved reducering af antallet af variable få en "dårligere" model - man må så afgøre, om denne models forklaringskraft er signifikant dårligere end den gamles.

Problemet om signifikans afgøres v.hj.a. kvotionettestet, som er 2 * differensen mellem de to log likelihood værdier fra de to modeller, der sammenlignes. Denne teststørrelse er chi-i-anden fordelt med k-n frihedsgrad, hvor k og n er antallet af parametre i hhv. den

---

(1): Udtrykket koefficienten/S.E. er en teststørrelse for hypotesen om, at det tilsvarende β er 0, når alle øvrige variable er inde i modellen. Udtrykket er normalfordelt, hvilket tillader, at der optræder negative værdier.

(2): Testet ser således ud: $\mathbf{U}'(0)\ I^{-1}\ \mathbf{U}(0)$, hvor $\mathbf{U}(0)$ er vektoren med de første afledte, med $\beta = 0$ indsat. I(0) er (minus den anden afledede), også kaldet informationsmatricen, med indsat $\beta = 0$.

nye og den gamle model. k-n vil typisk være lig 1, da vi i de fleste tilfælde kun undersøger én ny variabels betydning af gangen.

e. Vekselvirkninger.

Med 32 parametre i udgangspunktet er det en tvivlsom sag at arbejde for en yderligere udvidelse af modellen med nye parametre. Det er vi imidlertid nødt til, da de 32 variable forventes at indgå korreleret med hinanden - og da vekselvirkninger må forventes at eksistere på kryds og tværs i materialet.

På forhånd (i kapitel 6) har vi fundet frem til, at fire ideer til forbedringer af modellen burde undersøges.

Vi erstattede alderen ved studiets start med log(alder) og løste et nyt sæt likelihood ligninger og fik log likelihood værdien -7227,2221, som er mindre end udgangspnktet, og derfor forkaster vi dén ændring.

Vi ville have øget "køn"s forklaringskraft ved på den ene side at kombinere køn og (teksam+forvaltning), der forventedes at have mænd som de hurtigste, og på den anden køn og (resten af OB-fagene), hvor kvinder forventedes at være hurtigst. Men da det ville udvide modellen med 22 nye parametre (og det skulle gøres trinvist: 60 kr. pr. kørsel), opgav vi at teste denne vekselvirkning.

Omkring vekselvirkning mellem alder og natbasis fik vi naturligvis en forbedring, og ved denne udvidelse af modellen fik vi da log likelihood værdien til -7226,5768 og kvotienttest-størrelsen til 1,14. Dette er ikke nok til, at vi kan betragte denne model som bedre end udgangsmodellen.

Alder og sambasis gav -7227,1415 hvilket kun er meget

lidt bedre end udgangsmodellen.

Så tænkte vi, at alder og hver enkelt af de tre basisvariable måske gav et godt resultat. De 3 nye variables indførsel betød imidlertid, at log likelihood konvergerede mod et tal; men at beta'et for EJBASIS (variabel nr. 6) ikke ville konvergere (efter de 5 foreskrevne trin). Vi burde nok have taget én af gangen, men tvivler nu på, at det havde forbedret modellen signifikant.

Vi foretog dernæst en prøve på sammenhængen mellem alder og en række humanistiske OB-faggrupper, idet vi kunne antage, at ældre studerende måske klarede sig bedre hér end andre steder. Desværre lavede BMDP det samme nummer som før (da man skal tage én af gangen), så vi valgte den bedste ud (dén der ikke var nævnt i fejlmeddelelsen om ikke-konvergens).

Alder og dansk-historie viste sig at øge log likelihood med 0,0724, hvilket er alt for lidt til at betyde noget.

Alt i alt var det altså svært at få proppet nye forklarende variable ind på dette stade - hvilket helt sikkert er betinget af de mange eksisterende, der dækker forklaringerne pænt ind - i hvertfald indenfor de "kendte" variable. Det er klart, at personens indkomst, status, bopæl osv. stadig ville have kunnet forbedre modellens forklaringskraft (log likelihood værdi) - men det kan vi imidlertid ikke stille noget op ved. Men da vi senere ved eliminationsproceduren ikke vil glemme de én gang borteliminerede uafhængige variable (fordi de efter at andre variable er fjernet pludselig igen kan få betydning), vil vi heller ikke glemme at prøve den (eller de) bedste vekselvirkninger igen, når vi står med den endelige model.

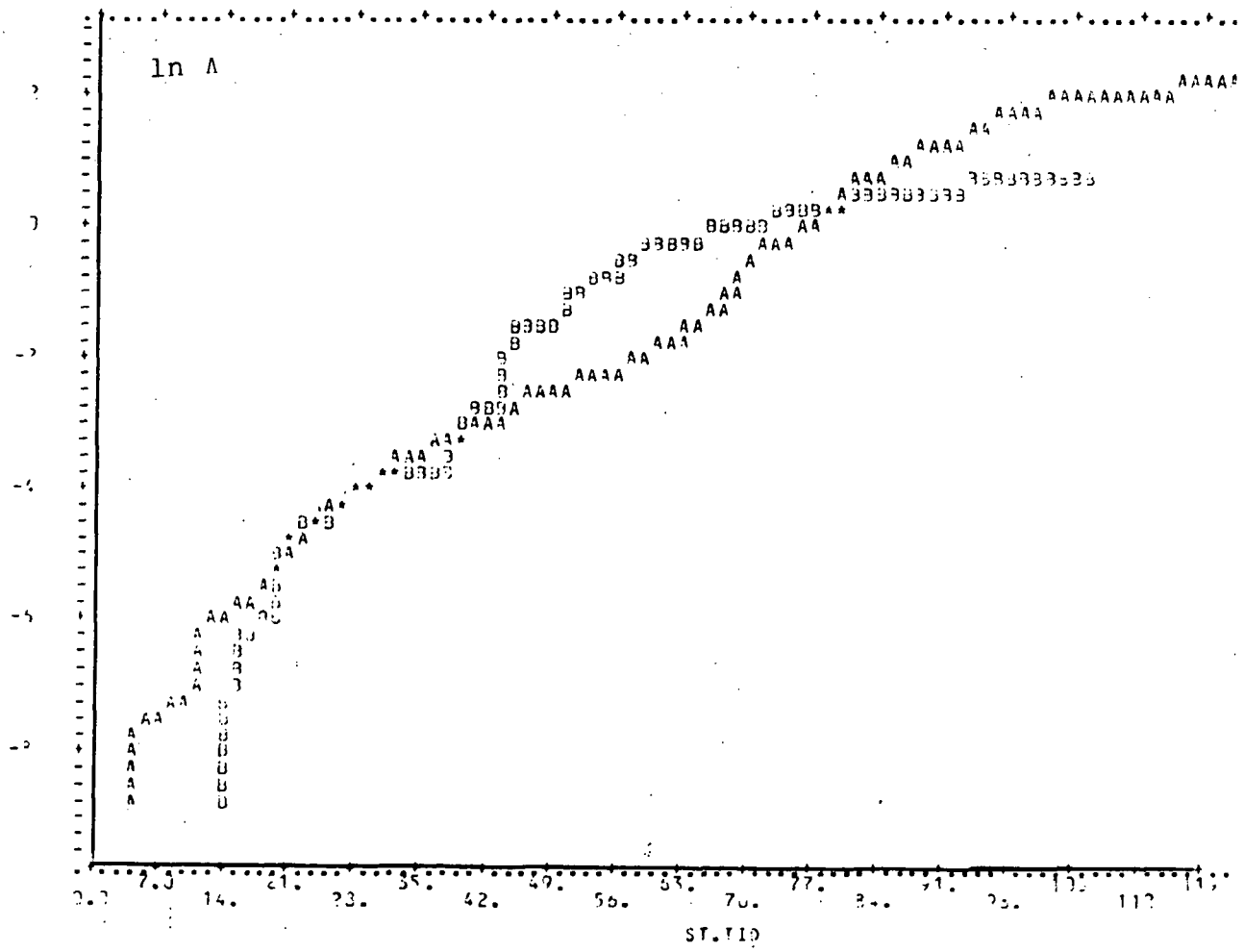
Behandlingen af vekselvirkningerne er helt igennem noget speget. Selvom vi nu tilgodeser dem inden eliminationsproceduren og også efter opstillingen af den

endelige model, kunne det jo godt være, at en vekselvirkning ville have kunnet få plads i modellen midtvejs i eliminationsproceduren - i fald vekselvirkningen pludselig fik signifikant betydning. (Da den forward selektion, vi benyttede, imidlertid opererede med meget sikre og rigelige marginer, og da chek'et bagefter var klart usignifikant, er vi nu ikke nervøse for vekselvirkningernes evt. undertrykkelse i vores konkrete modelopbygning).

f. Proportionalitet.

De næste skridt bestod i at undersøge proportionalitetsantagelsen i Cox-modellen. For køn og alders ved-

Figur 7.4 Proportionalitetstest af socionomi (B) mod de øvrige fag (A).



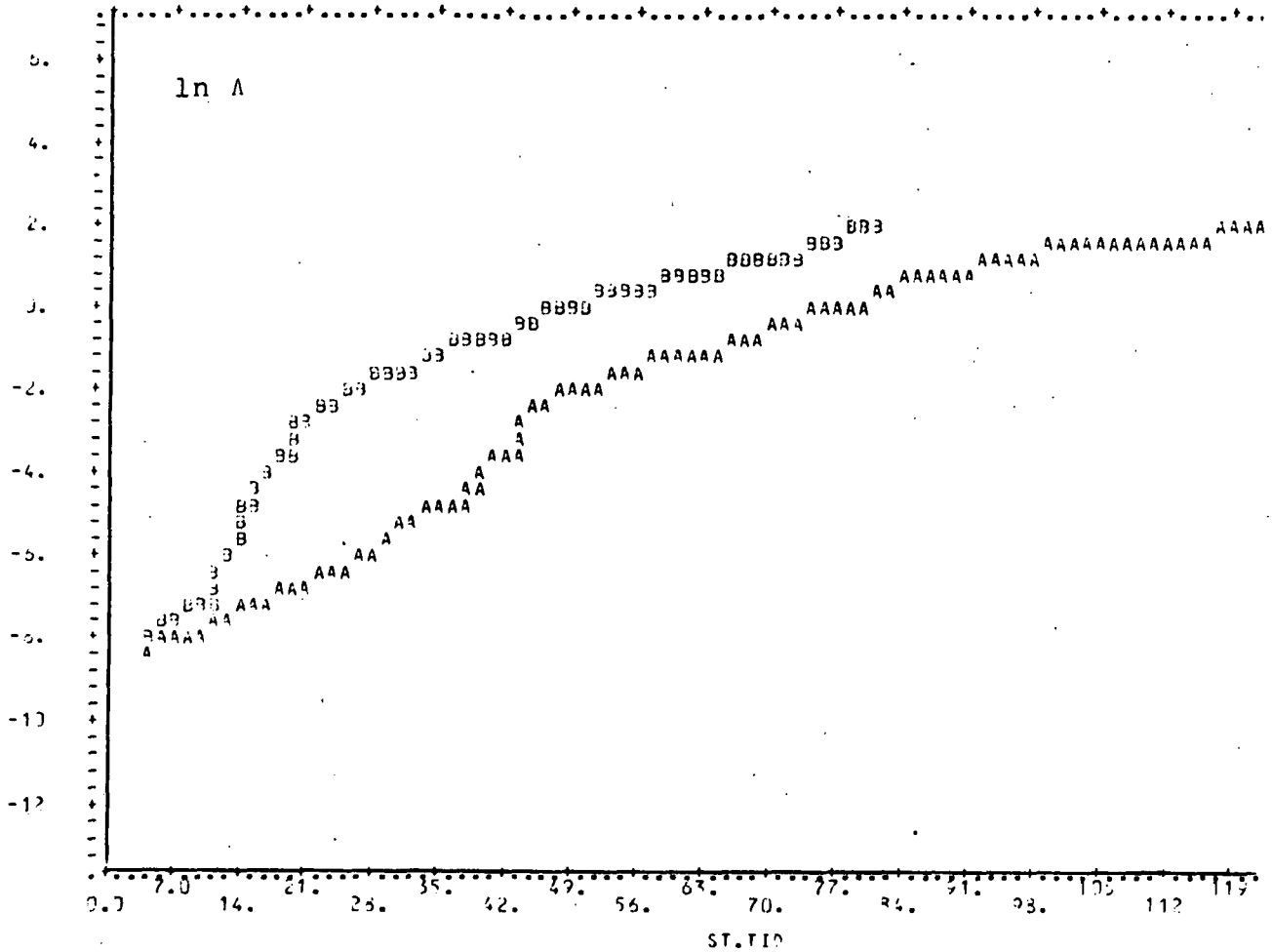
kommende var der ingen problemer (se i bilag D). For forvaltnings vedkommende var vi på forhånd mere spændte, da vi vidste, at nogen forlader RUC med den korte forvaltningsgrad, andre med kandidatgraden. Alligevel viste kurverne, at proportionaliteten var god nok (se bilaget).

Imidlertid gik det galt, da vi prøvede socionomi op mod de andre fag (se figur 7.4 på forrige side).

På figuren ser vi, at kurve B (socionomi) er væsentlig mere stejl i starten end kurve A (alle øvrige). De to kurver kan ikke siges at have samme forløb, selvom en numerisk test muligvis ikke kunne have forkastet proportionalitetsantagelsen. Vores interesse i at have socionomi inde i modellen er nu heller ikke så stor, da faget ikke længere eksisterer på RUC. Så med figuren som argument vil vi senere ændre udgangsmodellen, så alle personer med socionomi udgår af undersøgelsen. Vi kan ikke bare nøjes med at se bort fra parameteren "socionomi", fordi det vi har set på figuren er, at socionomi-personernes dødelighedsforløb i sin helhed er forskellig (dvs. ikke-proportional) fra resten.

Vi går nu videre i rækken af tests af proportionalitet. Denne figur (se næste side) skelner imellem "ejbasis" og resten. I begyndelsen af det interval, vi med rimelighed kan regne på, er den lodrette afstand 4 enheder, mens den er 1,5 ved afslutningen og falder jævnt herimellem. Dette er skarpt i modstrid med antagelsen om proportionalitet. Imidlertid er den vandrette afstand imellem kurverne umiskendelig tæt på 24 mdr. i hele det relevante interval. Dette kunne berettige til at prøve at forskyde kurven for ejbasis-folk med 24 mdr. mod højre, hvilket falder fint i tråd med, at man netop springer de 24 mdr. over, basisuddannelsen tager for de øvrige personer. Dette vil vi vende tilbage til senere.

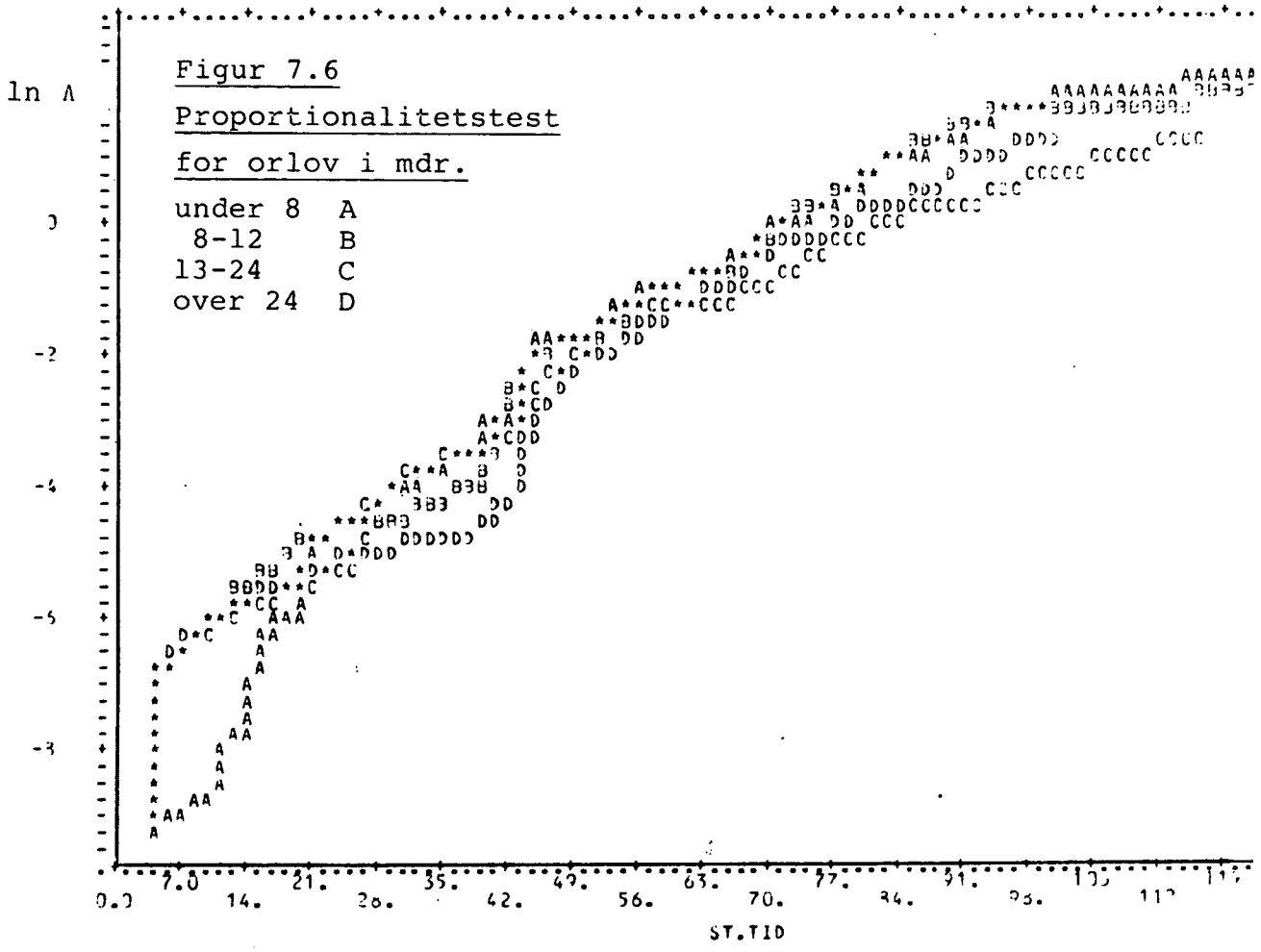
Figur 7.5 Proportionalitetstest af ej basis (B) mod resten (A).



Hvad angår orlovtid og studieskift, er der ingen tvivl om, at der gælder proportionalitet.

Se figurene 7.6 og 7.7 på næste side.

Men til og med kunne man få en mistanke om, at det ikke var særlig nødvendigt at inddrage oplysninger om deres kvantitative vægt, men måske blot dét, om der har været tale om orlov/studieskift eller ej. Også dette vil vi vende tilbage til senere.

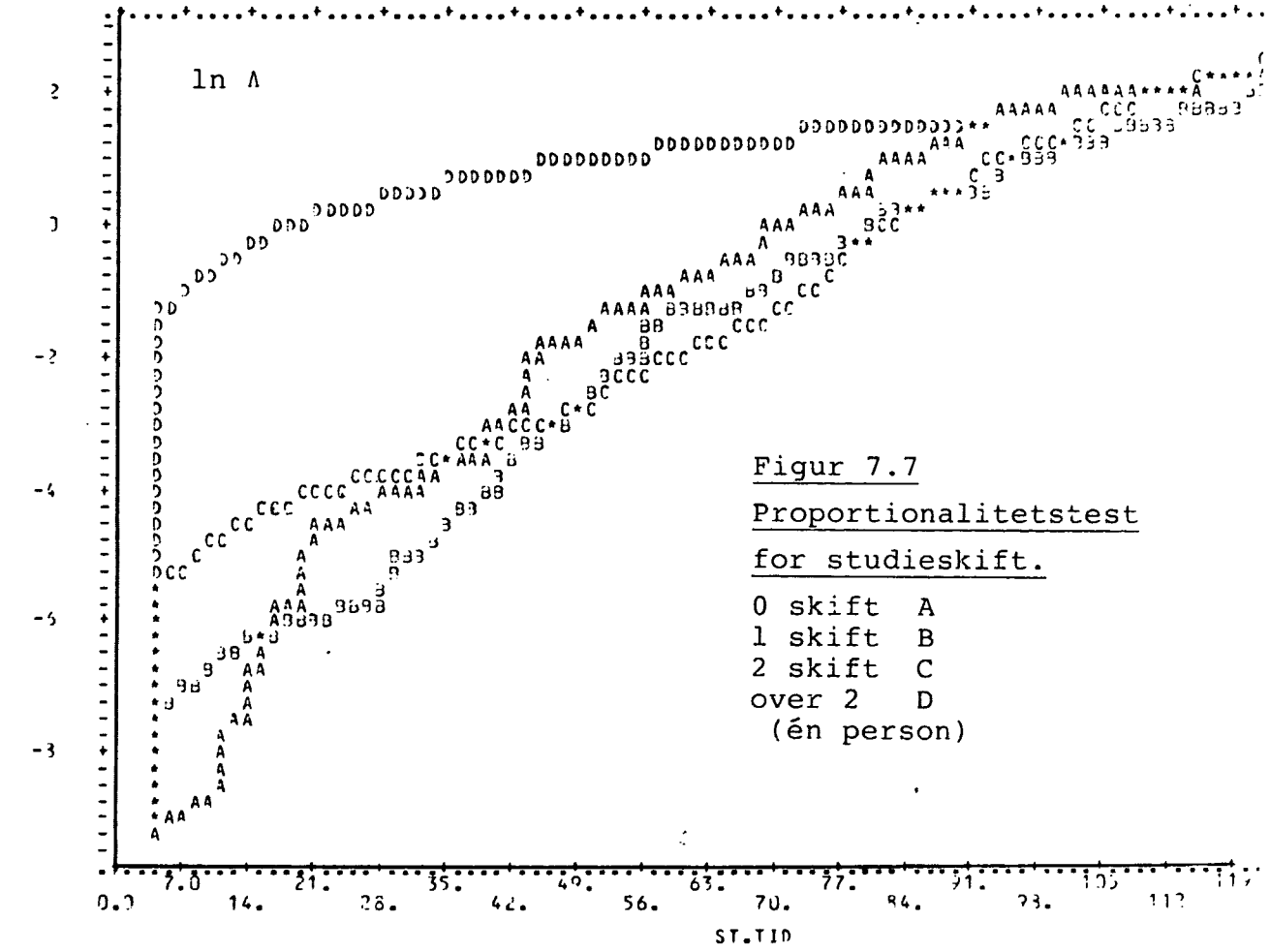


Figur 7.6

Proportionalitetstest

for orlov i mdr.

under 8 A
8-12 B
13-24 C
over 24 D



Figur 7.7

Proportionalitetstest

for studieskift.

0 skift A
1 skift B
2 skift C
over 2 D
(én person)

Figur 7.8 Koefficienterne i den anden fulde model.

LOG LIKELIHOOD = -3462.5476
GLOBAL CHI-SQUARE = 411.61 D.F.= 31 P-VALUE = .0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
3	KØN	.0091	.0934	.0974	1.0091
4	STARTAAR	.0284	.0255	1.1161	1.0288
5	STARTALD	-.0206	.0103	-1.9977	.9796
6	EJBASIS	-.2114	.2823	-.7491	.8094
7	SAMBASIS	.0704	.1335	.5275	1.0729
8	MATBASIS	.3619	.1815	2.1041	1.4651
9	BASISTID	-.5572	.0999	-5.8747	.5559
10	ST.SKIFT	-.3356	.1016	-3.2996	.7152
11	ORLOVTID	-.0289	.0071	-4.0899	.9715
12	BIOLOGI	-3.2052	.6021	-5.3233	.0405
13	BIOKEMI	-3.2848	.6474	-5.0737	.0374
14	DANENG	-3.0343	.6435	-4.7156	.0431
15	DANHIST	-3.3243	.5722	-5.8097	.0360
16	DANSAM	-2.9137	.5905	-4.9345	.0543
17	MATFYS	-3.8210	.6718	-5.6874	.0219
18	GEOHIST	-3.3116	.5867	-5.6444	.0365
19	GEOSAM	-2.8591	.5561	-5.1415	.0573
20	HISTSAM	-3.2315	.5513	-5.8618	.0395
21	GEOGRAFI	-.5434	.6767	-.8015	.5814
22	PSYKOLOG	-1.0689	.5959	-1.7939	.3434
23	HISTTYSK	-2.7623	.6563	-4.2099	.0631
24	TEKSAM	-3.5261	.5385	-4.5215	.0295
25	SOCMEDIE	-.5869	.5532	-1.0515	.5562
26	FORVALTN	-2.7861	.5420	-5.1036	.0617
27	NAT	-1.3893	.7706	-1.8029	.2492
28	HUM	-2.0382	.7792	-2.6159	.1303
29	NATNAT	-3.7122	.6322	-5.8714	.0244
30	HUMHUM	-3.2355	.5569	-5.8099	.0393
31	SAMNAT	-2.9804	.5503	-5.3512	.0187
32	SAMHUM	-3.2265	.5610	-5.7509	.0327
37	NATHUM	-3.6037	.6148	-5.8701	.0271

g. Omformulering af modellen.

På baggrund af erfaringerne med proportionalitetstest'
ene måtte vi ændre på udgangsmodellen (socionomi
er fjernet og der er lagt 24 mdr. på hos ej-basisterne).

Den nye likelihood værdi er større, fordi der er væsentlig færre med i modellen. Den mest markante forskel fra den tidligere "fulde" model er naturligvis parameter-værdien udfor ejbasis. Noget overraskende er værdien nu gået hen og er blevet negativ (relativt mindre døds-intensitet og dermed længere studietid til eksamen); men hér må vi huske på, at vi snakker om relativ

dødsintensitet i forhold til en (ukendt) referenceperson, der har en z-vektor = **0**-vektor. Personen er f.eks. 0 år gammel, startede i år 1900, går på humbasis og bruger 0 år på at bestå basis. OB-faget kan man tænke på som et af dem, der ikke er nævnt (ex: driftsøkonomi). Beta-værdien siger altså især noget set i forhold til de øvrige beta-værdier. Den absolutte værdi har betydning, hvis vi har et rimeligt kendskab til referencepersonens udseende.

Vi kunne have skabt os en mere passende referenceperson ved f.eks. at bruge befolkningens gennemsnits-værdier som en faktor, der automatisk trækkes fra de kvantitative variable (ex: startalder: z_i - 22 år). Evt. kan man nøjes med at trække den mindst fundne z-værdi fra (72 og 17for hhv. startår og startalder, se den deskriptive statistik, figur 7.2) for at undgå negative z-værdier efter subtraktionerne. En sådan reference (gennemsnitsperson) vælger BMDP, når den præsenterer overlevelsesfunktionens graf (se figur 7.14), for at bibringe et indtryk af den afbillede persons placering i "feltet" af studerende. Vi har alligevel valgt at lade være, fordi det let vil kunne øge forvirringen, især omkring parametrenes relative betydning.

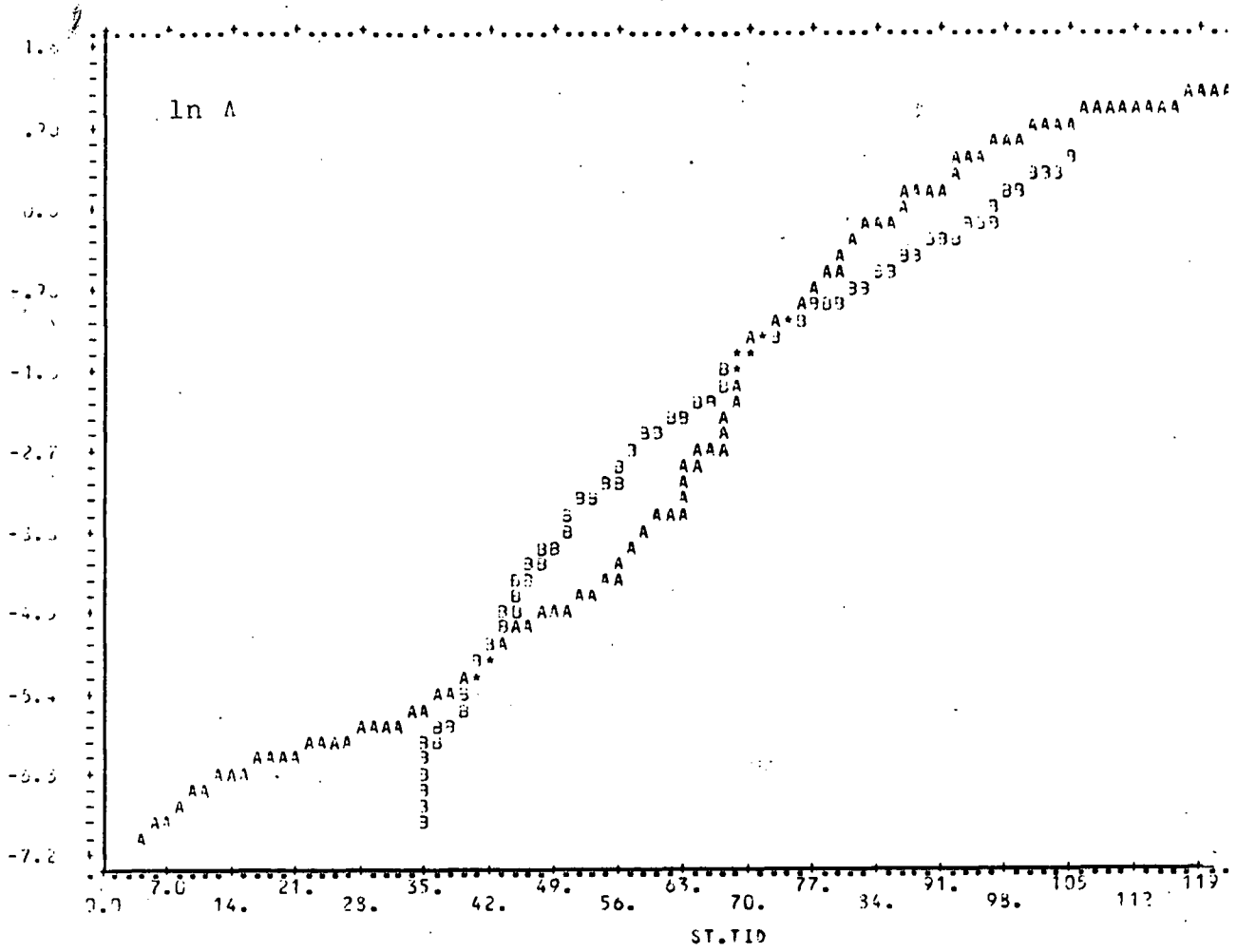
Ved at lægge 24 mdr. til hver enkelt persons studietid, når der er tale om folk uden basis, paralelforskyder vi $\log\Lambda_B$-kurven hen over $\log\Lambda_A$-kurven.

Se figur 7.9 på næste side.

Som det ses, kommer "ejbasis"-kurven til at opføre sig nogenlunde som socionomi's kurve tidligere. Det er nok heller ikke så mærkeligt, da de fleste uden basis kun skal læse 3 moduler (ét fag) på OB ligesom socionomerne skulle.

Hvor gerne vi end ville undgå det, tyder alt på, at modellen bør opdeles; én for folk uden basis og én

Figur 7.9 Proportionalitetstest af ny ej basis (B) og øvrige (A).



for folk med basis. Da vores interesse for folk uden basis ikke er så stor i denne sammenhæng (det er en meget uhomogen gruppe, der ikke nødvendigvis beskrives særlig godt med vores variable), vil vi vælge blot at fjerne disse personer fra vores videre undersøgelser.

Vi undersøgte på dette stade, om man kunne forenkle modellen ved at ændre orlovstiden til en 0-1 variabel; altså blot skelne mellem om man har haft orlov eller ej. Men denne mere simple model gav en log likelihood værdi på -3465,5938 hvilket afslører en signifikant dårligere model.

Derimod viste det sig, at en model med studieskift som 0-1 variabel var indenfor den grænse, vi ville acceptere for at tro, at forværringen i modellens forklaringskraft var tilfældig. Log likelihood blev hér

-3461,1245, og kvotienttestet holder sig indenfor 10% niveauet, der er det niveau, der benyttes i BMDP modellen.

h. Ny fuld model.

Efter nu at have undersøgt modellen for vekselvirkninger, undersøgt proportionalitet (og herunder smidt en del af befolkningen bort!) og forenklet modellen (enkelt-OB-fag er f.eks. væk fra regressionen) står vi nu overfor den færdigbearbejdede udgangsmodel.

26 variable er tilbage og 1039 personer danner materialet bag undersøgelsen, heraf er 50% censurerede (aktive studerende pr. 1/9 1984).

Figur 7.10 Koefficienterne i den sidste fulde model.

LOG LIKELIHOOD = -2852.1392
GLOBAL CHI-SQUARE = 174.06 D.F.= 26 P-VALUE = .0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
3	KØN	.0267	.1033	.2586	1.027
4	STARTAAR	.0056	.0290	.1996	1.006
5	STARTALD	.0035	.0110	.3208	1.004
7	GAMBASIS	.1266	.1425	.8864	1.135
8	MATBASIS	.4882	.1725	2.5307	1.629
9	BASISTID	-.6000	.0990	-6.1495	.549
10	ST.SKIFT	-.4497	.1222	-3.6803	.638
11	UDLOVTID	-.0276	.0075	-3.6430	.973
12	SIDSEC	-2.0442	.6875	-2.9734	.129
13	BIOKEMI	-2.0312	.7432	-2.7330	.131
14	DAMENG	-1.7494	.7161	-2.4416	.174
15	DANHIST	-1.2410	.6565	-2.9575	.289
16	DANSAM	-1.7208	.6701	-2.5677	.179
17	MATFYS	-2.6252	.7502	-3.4996	.072
18	GEOHIST	-2.1743	.6885	-3.1582	.113
19	GEOSAM	-1.5954	.6474	-2.4645	.203
20	HISTSAM	-1.9922	.6307	-3.1141	.136
23	HISTYSK	-1.5443	.7026	-2.0068	.213
25	TEKSAM	-1.1950	.6319	-1.8912	.302
26	SOCMEDIE	.6702	.7091	.9461	1.955
30	FORVALTN	-1.5337	.6341	-2.4658	.216
32	MATMAT	-2.9115	.7418	-3.7839	.054
34	HUMHUM	-1.8471	.6407	-2.8828	.157
35	SAMMAT	-2.9262	.7626	-3.8374	.053
36	SAMHUM	-1.9384	.6444	-3.0079	.143
37	MATHUM	-2.2963	.7019	-3.2716	.100

Vi bemærker, at basistiden har størst forklaringskraft (når alle 26 er inde som ligestillede parametre), der-

efter følger samnat, natnat, studieskift, orlovtid, matfys osv.

En varians/covarians matrix giver et broget billede af de forskellige sammenhænge i de 26 resterende variable. Her findes f.eks. en negativ covarians mellem koefficienterne til natbasis og natnat på -0,0274 med den tilsvarende korrelationskvotient på -0,19, henover værdier på 0 (køn og startalder er f.eks. slet ikke korreleret) til korrelationsværdier tæt ved + 1. (1)

For at gøre den endelige model mere simpel skal vi nu til at eliminere variable. Ækvivalent hermed forventer vi, at korrelationerne vil nærme sig 0 for de bedste parametre i regressionen (blandt dem, vi vil "nøjes" med at benytte).

i. Opbygning af den endelige model.

BMDP har indbygget flere procedurer for dette. Der er rig mulighed for selv at styre processerne. Vi har valgt at lave en forward selektion (udvælgelse), hvor BMDP benytter kvotienttestet til at beregne forbedringer/forværringer i modellens forklaringskraft. Fremgangsmåden hedder MPLR (Maximum Partial Likelihood Ratio test). Der fastsættes nogle grænser for, hvor lille en forbedring og hvor lille en forværring, der

---

(1): $\text{Korrelation}(A,B) = \dfrac{\text{covarians}(A,B)}{\text{S.E.}(A)\cdot\text{S.E.}(B)}$

Covariansen fås udfra informationsmatricen $-(l''(\beta))$, nemlig "Varians-covarians matrice" = $-(l''(\beta))^{-1}$, hvor diagonalen er varianserne og de øvrige covarianserne (spejlet om diagonalen). Standardafvigelsen fås ved at tage kvadratroden af covariansen.

må accepteres for hvert tænkeligt skridt (10%-niveau for at gå ind og 15% for at gå ud).

Vi lod alle variable starte uden for modellen, da vi ikke har kunnet pege på én eller flere variable, der "erfaringsmæssigt har betydning".

MPLR-proceduren gik lige frem til sit mål. Den brugte 7 skridt til at putte 7 variable ind i modellen, og på dette trin var der ikke flere variable, der kunne flyttes (som berammet af de fastsatte signifikans-niveauer).

Figur 7.11 Koefficienterne i den endelige model.

LOG LIKELIHOOD = -2863.1896
IMPROVEMENT CHI-SQ (2*(LN(MPLR))) = 5.77 D.F.= 1
GLOBAL CHI-SQUARE = 149.67 D.F.= 7

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
4	BASISTID	-.5571	.0958	-5.8106	.5728
10	ST.SKIFT	-.4513	.1162	-3.8841	.5366
11	ORLOVTID	-.3240	.0972	-3.6349	.7233
19	GEOSAM	.4536	.1579	2.5245	1.5737
21	TEKSAM	.2136	.1109	3.2575	2.4377
28	SOCMEDIE	2.5903	.3995	6.4535	13.3511
30	FORVALTN	.4457	.1363	3.2704	1.5614

Her ser vi den endelige model. De 7 variable er placeret i [illegible] og beta-værdier osv. De kom ind i følgende rækkefølge:

Figur 7.12 Opsamling af trinene i udvælgelsesproceduren.

STEP NO	VARIABLE ENTERED	DF	LOG LIKELIHOOD	IMPROVEMENT CHI-SQUARE	P-VALUE	GLOBAL CHI-SQUARE
0			-2928.775			
1	21 TEKSAM	1	-2911.211	35.134	.000	40.289
2	4 BASISTID	2	-2896.604	29.214	.000	69.503
3	10 ST.SKIFT	3	-2886.100	21.008	.000	90.511
4	28 SOCMEDIE	4	-2877.534	17.132	.000	107.643
5	11 ORLOVTID	5	-2870.105	14.863	.000	122.506
6	30 FORVALTN	6	-2868.079	4.053	.044	126.559
7	19 GEOSAM	7	-2863.190	5.773	.016	132.331

Differencerne ml. hvert trins log likelihood værdier ganges med to og kaldes forbedring i chi-i-anden (det er kvotienttesstørrelsen). Den næste forbedring (trin 8) ville i så fald have været indsættelsen af samnat, der med sin chi-i-anden værdi på 2,53 lige netop ikke accepteres på 10%-niveauet.

Det generelle indtryk af varians/covarians matricen

er, som forventet, mere ensartet og tættere ved 0. Den største covarians findes mellem teksam og geosam, hvor værdien 0,0041 kan regnes om til en korrelationskvotient på 0,22. Til sammenligning var denne korrelation inden eliminationsproceduren 0.96. Overlappet er så at sige forsvundet ved at eliminere variable.

Figur 7.13 Korrelationsskema ml. de variable i den endelige model.

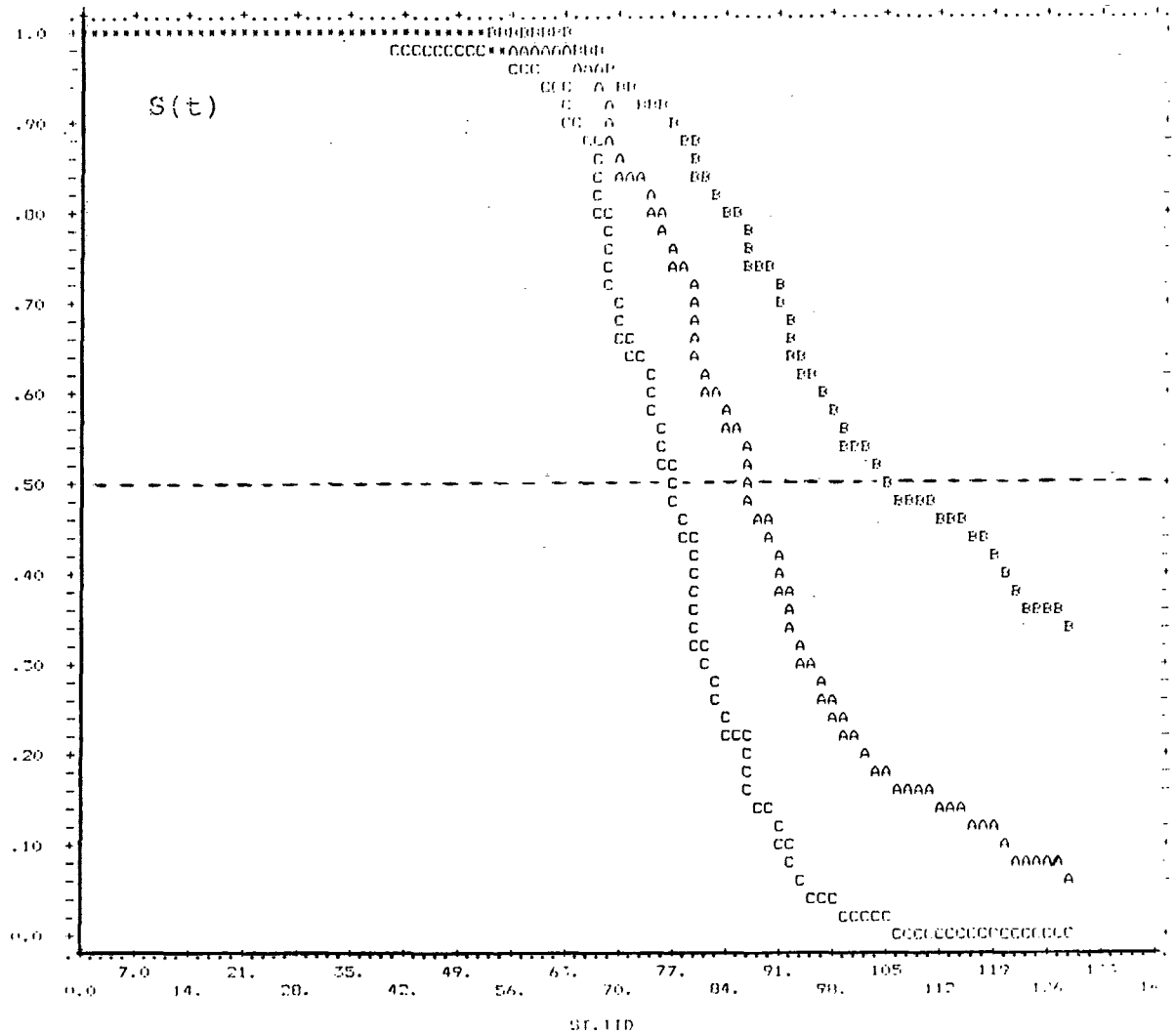
	Basis	St.skift	Orlovtid	GeoSam	Teksam	Socmedie
St.skift	0,0683					
Orlovtid	0,1305	-0,0717				
GeoSam	-0,0074	-0,0649	0,0662			
Teksam	-0,2180	-0,0342	-0,0377	0,2208		
Socmedie	-0,0319	0,0513	-0,2121	0,0495	0,1258	
Forvaltn	-0,0146	0,0625	-0,0815	0,1695	0,2740	0,1045

Den største korrelation viser sig at være imellem teksam og forvaltning (0,2740). Dvs. at dé to variable overlapper hinanden mere end nogle af de andre gør. I større omfang (numerisk større tal) ville det kunne have betydet, at den mindst signifikante variabel af dé to ville udgå, idet den blev "forklaret" af den anden.

Det bemærkes, at beta-værdierne i den endelige model generelt er blevet mere positive nu, end de var inden eliminationen. Det dækker over, at "referencepersonen" har ændret sig (refererer til et længere levende individ). Basistiden, studieskift og orlovtiden vil for øgede værdier øge levetiden på RUC; mens de øvrige vil forkorte levetiden. De 4 fag(-kombinationer) er altså markant medvirkende til at give hurtigere eksamen. For at illustrere dette har vi fået udtegnet 3

personers overlevelseskurver.

Figur 7.14 Overlevelseskurver til eksamen for 3 udvalgte personer.



PATTERN	9 BASISTID	10 ST.SKIFT	11 ORLOVTID	19 GEOSAM	26 TEKSAM	28 SOCMEDIE	30 FORVALTN
1	2	0	0	0	0	0	0
2	3	1	12	1	0	0	0
3	2	0	0	0	1	0	0

PATTERN	SYMBOL	CONVERSION FACTOR **
1	A	.893
2	B	.354
3	C	2.227

Omregningsfaktoren, der omtales, beskriver personernes overlevelse set i forhold til en tænkt gennemsnitsperson.

Som det fremgår, har vi søgt at ramme en almindelig gymnasielærer-RUC'er (A), en god studenterpolitisk og samfundsengageret RUC'er (B) og en stræbsom RUC'er (C). Kurverne når ikke alle ned til 0, da der øjensynlig befinder sig studerende på A og B med længere studietider end 130 netto-måneder.

Medianen er indtegnet og udtrykker, at halvdelen af personerne med pågældende variable efter det viste antal måneder nu har opnået eksamen.

Det fremgår ikke direkte af sammenhængen hvor mange personer, der "gemmer" sig bagved hver enkelt kurve. Det skal - om nødvendigt - undersøges direkte fra data-basen.

På trods af, at teksam på linie med de øvrige fag(-kombinationer) er en kandidatuddannelse, er den den kraftigste faktor, der markerer et hurtigt studieforløb til eksamen.

Vi skal ikke forsøge os med de "bagkloge" forklaringer - vi mener nemlig, at tallene skal give anledning til debat fremfor entydige forklaringer inden der er foretaget yderligere undersøgelser med andre midler end denne metode.

Alligevel må teksam's mere homogene studieforløb (alle 6 moduler samme sted, under samme planlægning) være en væsentlig del af forklaringen.

Variablen socmedie dækker over 8 personer, i det generelle billede er det ikke en særlig vigtig oplysning, som modellen bidrager med hér. Den er med, fordi det er en 2 * 3 modulers OB-uddannelse, der var utraditionel og langt hen ad vejen noget på tværs på RUC. Alligevel klarede de 8 sig fint med denne kombination.

At basistiden vejer så tungt, overraskede os noget.

Som før nævnt er denne variabel lidt tvivlsom, idet den pr. automatik er positivt korreleret med studietiden. Men det er vores erfaring udfra vort kendskab til data-basen, at det specielt er nogle få personers lange orlovsperioder i basis, der skaber de helt lange basistider, samt et større antal studietidsforlængelser i basis på hvert ½ år, der af tekniske grunde betyder, at basistiden i så fald rundes op til 3 år.

Korrelationen mellem basistid og orlovstid er overraskende lille, nemlig +0,11.

Den omstændighed, at man har haft studieskift, og længden af orlovstiden, vejer i sig selv ret tungt i retning af en studietidsforlængelse inden eksamen. Bemærk at selve orlovsperioden selvfølgelig er trukket fra i studietiden - men orloven giver altså ikke friske kræfter til fornyede studier.

Geosam var så den eneste gymnasielærer kombination, der formåede at komme ind i modellen. Ligesom ved teksam, vil det nok ikke være forkert at pege på et relativt homogent og integreret studie-miljø som en forklaringsfaktor for det hurtigere forløb.

j. Afgangsmodellen.

Vi valgte nu at bruge de samme variable som ved eksamensmodellen, dog incl. studietrin og excl. socmedie, der kun rummer eksaminer. Vi havde simpelt hen ikke bevillinger til mere - så selvom i hvertfald proportionalitetsantagelserne er lidt tvivlsomme at overføre direkte - var det dog bedre at bruge de sidste penge på dette spændende emne, forbeholdene inkluderet, end at lade være.

Den fulde model med 26 variable inde ser således ud:

Figur 7.15 Koefficienterne i den fulde afgangsmodel.

LOG LIKELIHOOD = -19225.4895
GLOBAL CHI-SQUARE = 2772.04 D.F.= 26 P-VALUE = .0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
3	KON	.2595	.0396	6.5566	1.2962
4	STARTAAR	-.0799	.0053	-15.1307	.9232
5	STARTALD	-.0141	.0037	-3.8084	.9860
7	SAMBASIS	-.0633	.0439	-1.4404	.9387
8	NATBASIS	.0326	.0548	.5940	1.0331
9	STUDTRIN	-.8242	.0443	-18.6120	.4386
10	BASISTID	-.2275	.0307	-7.4029	.7965
11	ST.SKIFT	-.6957	.1516	-4.5902	.4987
12	ORLOVTID	-.0120	.0046	-4.1076	.9812
13	BIOGEO	-.2659	.4376	-.6075	.7666
14	BIOKEMI	-.4538	.5173	-.8772	.6352
15	DANENG	-.1506	.3061	-.4920	.8602
16	DANHIST	-.1202	.2445	-.4916	.8867
17	DANSAM	.0240	.3010	.0798	1.0243
18	MATFYS	.3451	.5907	.4732	1.4122
19	GEOHIST	.1272	.2700	.4713	.8805
20	GEOSAM	-.3543	.2176	-1.6280	.7017
21	HISTSAM	-.0326	.1688	-.1930	.9679
24	HISTTYSK	.2449	.4276	.5727	1.2775
27	TEKSAM	-.5032	.1828	-2.7806	.6016
31	FORVALTN	.0731	.1748	.4181	1.0758
34	NATNAT	-.6751	.3256	-2.0735	.5091
35	HUMHUM	-.0309	.1321	-.2339	.9696
36	SAMNAT	-.7334	.4320	-1.6975	.4803
37	SAMHUM	-.5369	.2290	-2.3450	.5845
38	NATHUM	-.3868	.3746	-1.0326	.6792

Kun 6 variable ser ud til at forøge dødeligheden (dem med positiv koefficient); dvs. folk hopper fra tidligt i studiet i forhold til baggrundspersonen. Bemærk, at det store flertal hopper fra inden OB-fag vælges, så i denne undersøgelse vil vi klart nok få vægtet de ikke-OB-afhængige variable særligt tungt. (Baggrundspersonen har ingen OB-fag og 0 i studietrin, det er derfor ikke underligt at OB-fagene generelt forlænger studietiderne).

Den endelige model fik vi frem på samme måde som før, dog afbrød vi regnearbejdet efter 10 minuters forløb. EDB-maskinen havde på de 10 min. nået at gennemføre 7 trin og indført 7 variable i modellen:

Figur 7.16 Koefficienterne i den endelige afgangsmodel.

```
                                LOG LIKELIHOOD =  -19240.8047
IMPROVEMENT CHI-SQ  ( 2*(LN(MPLR) ) =      15.60   D.F.=    1
                    GLOBAL CHI-SQUARE =    2733.32   D.F.=    7

                                 STANDARD
   VARIABLE  COEFFICIENT          ERROR      COEFF./S.E.   EXP(COEFF.)
   --------  -----------        --------     -----------   -----------
 3 KØN            .2527          .0390         6.4843        1.2875
 4 STARTAAR      -.0814          .0052       -15.5362         .9218
 5 STARTALD      -.0141          .0037        -3.8514         .9860
 9 STUDTRIN      -.8631          .0361       -23.9090         .4219
10 BASISTID      -.2275          .0308        -7.3808         .7965
11 ST.SKIFT      -.7335          .1494        -5.2783         .4545
12 ORLOVTID      -.0187          .0046        -4.0424         .9814
```

Udover de 7 udvalgte, ser det ud til at yderligere tre variable ville kunne komme med i den endelige model; nemlig teksam, samhum og sambasis, som alle vil forlænge studietiden.

Kun den omstændighed, at man er kvinde forkorter studietiden indtil en afgang uden eksamen.

Ved senere startår på RUC er der en tilbøjelighed til at forlænge sine studier i modsætning til de tidlige årgange på RUC.

Også øget alder ved immatrikulationen betyder forlængede studier i forhold til en lav alder.

At øget studietrin forøger læsetiden er vel ikke så underligt. Man får altså også bestil noget mens man "afventer" sin afgang uden eksamen.

Øget basistid, studieskift og orlov forlænger meget rimeligt studietiden, ligesom i eksamensmodellen.

8. KONKLUSION.

Vi vil i denne konklusion ikke focusere - som det måske kunne forventes - på output'et af vor undersøgelse. Grunden til dette er dels, at vi finder det for kategorisk at udtale os om studiemønstre ud fra denne undersøgelse - dels pga., at vort formål med projektet var selve arbejdet med modelbygningen. Vi vil derfor i denne konklusion koncentrere os om selve processen i projektet.

Vi har formået, fra bunden af, at skaffe os et overblik over statistiske overlevelsesmodeller og blandt dem især Cox's regressionsmodel. Vi oparbejdede et så godt kendskab til denne model, at vi kunne forholde os så tilpas kritisk og skabende til den, at vi synes, at vores egen model-opstilling hviler på et forsvarligt teoretisk grundlag. Godt nok fik vi et alvorligt grundstød, da vi ramlede ind i problematikken omkring de konkurrerende dødsårsager; men vi reddede os helskindede igennem. I den forbindelse måtte vi til lejligheden lave vor egen censureringsprocedure, hvor vi fordelte censureringerne på de to dødsårsager.

Vi har fået et praktisk kendskab til problemer i forbindelse med indsamling og bearbejdning af store datamængder, fået kendskab til opbygning af en database, og har fået et mere nærværende forhold til datalogien som hjælpemiddel i matematikken.
Vore oplevelser med datakørsler på RECKU - og specielt priserne på disse kørsler - tvang os til at prioritere vort arbejde, således at ikke alle kørsler, der oprindeligt var planlagt, blev gennmført.

Det er rent faktisk lykkedes at opstille en model for, hvordan eksaminerne fordeler sig over tiden afhængig af en lang række personlige variable - og en tilsvarende model for folk, der forlader RUC uden eksamen, er

lavet. Hermed har vi vist, at Cox-modellens anvendelsesområder ikke blot ligger på lægevidenskabelige og biologiske felter.

Det er overordenligt vigtigt at stille de rigtige spørgsmål, når en sådan undersøgelse skal tilrettelægges. Vi har måske spændt os for vidt i vores undersøgelse for at få et generelt billede af RUC-studenternes studiemønstre. Godt nok blev vi tvunget til at fjerne større grupper af personer; men vi kunne måske godt have focuseret mere præcist på enten basis' problemer eller overbygningens problemer. Men ingen vidste noget særligt om forholdene inden vores undersøgelse, så det er nok urimeligt at bebrejde os noget i dén sammenhæng.

De indgående variable tåler gerne mere bearbejdning, så data-basen står til rådighed for nye penneførere. Variablene er ret uensartede, og især 'basistid' og 'studietrin' bør studeres nøjere.

Det var opsigtsvækkende, at kun få gymnasielærer-kombinationer kom med i eksamensmodellens endelige parametre, og det må siges, at mange spændende oplysninger er tilvejebragt i undersøgelsen. Det kan eksempelvis nævnes, at forudsat en person får eksamen med to gymnasielærerfag er det ret ligegyldigt, hvilke to fag, der studeres. Studietiden på den ene kombination er stort set den samme som på den anden.

Vi har skrevet et projekt, der forhåbenlig både i form og indhold kan formidle et spændende statistisk område til andre, og hér tænker vi specielt på andre studerende, der kunne tænkes at ville arbejde videre med stoffet.

Alligevel mangler der en oplagt model - der dog kan tilvejebringes. Denne model - Q-modellen - fortæller om en persons chancer for - efter et vist antal år med de for personen specielle baggrundsvariable - at

få en eksamen, at få afgang eller stadig at være aktiv studerende.

Udfra undersøgelsens umiddelbare relevans har vi skabt et samarbejde med RUC-administrationen og håber på at det vil fortsætte, og at administrationen måske kan drage nytte af undersøgelsen. Vi agter at skrive til RUC-NYT om resultaterne, så der kan etableres en egentlig debat, der måske kan stimulere interessen for yderligere undersøgelser, gerne med et helt andet udgangspunkt, end dét, vi har haft.

9. LITTERATURLISTE.

ANDERSEN, P.K.: Testning Goodness of Fit of Cox's regression and Life Model. Biometrics 38, 67-77 1982.

ANDERSEN,P.K.& RASMUSSEN,N.K.: Admission to Psychiatric Hospitals among Women Giving Birth and Women Having Induced Abortion. A Statistical Analysis of a Counting Process. Statistical Research Unit Research Report 6. 1982

ANDERSEN,P.K & VÆTH,M.: Statistisk analyse af overlevelsesdata ved lægevidenskabelige undersøgelser. FADL's forlag 1984.

BONNEVIE,O. et al.: Overlevelsesmodeller i klinisk forskning. Ugeskrift for læger nr.38 september 1971.

BRESLOW,N.E.: Analysis of Survival Data under the Proportional Hazards Model. Int. Stat. Rev., vol 43 no. 1,1975 pp 45 - 58.

CASPERSEN,L.C. et al.: Cox's regressionsmodel - anvendt på overlevelsesdata vedrørende maglingt melanom. Specialerapport. Aalborg Universitetscenter 1984.

CHRISTENSEN,E. et al.: A Therapeutic Index That Predicts the Individual Effects of Predrisone in patients with Cirrhosis. Gastroenterology 1985:88:156-165.

COX,D.R.: Regression Models and Life Tables. Journal of the Royal Statistical Society. B 34. 1972.

COX,D.R. & OAKES,D.: Analysis of Survival Data. Capman and Hall Ltd. 1983.

DIXON,W.J.: BMDP Statistical Software. University of California Press. 1983 (kap. 19)

GAIL,M.:A Review and Critique of Some Models Used in Competing Risk Analysis. Biometrics 31, 209-222 1975.

HARALDSSON,A.: Programmering i pascal. Teknisk Forlag a/s 1979.

JUEL,K. & JACOBSEN,P.: Dødeligheden 1956-80 i Thyborøn-Harboør og i et kontrolområde. Hygiejnemeddelelser 1984.

KALBFLEICH,J.D. & PRENTRICE,R.L.: The Statistical Analysis of Failure Time Data. John Wiley and Sons. 1980.

KALBFLEICH,J.D. & PRENTRICE,R.L.: Marginal Likelihoods Based on Cox's Regression and Life Model. Biometrica 1973, 60,2. pp 267-278.

PETO,R. et al.: Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient. I.Introduction and Design. Br.J.Cancer(1976) 34.585-610 and II. Analysis and Exaples. Br.J. Cancer(1977) 35. 1-39.

PRENTRICE,R.L. et al.: The Analysis of Failure Times in the Precence of Competing Risks. Biometrics 34, 541-554. 1978.

ROSKILDE UNIVERSITETSCENTER:Akademisk Forvaltning. Studiestatistik (opgjort pr. februar 1981) 1982.

SCHLICHTING,P. et al.:Prognostic Factors in Cirrhosis Identified by Cox's Regressions Model. Hepatology Vol.3.No.6. 889-895. 1983.

STATISTISK ÅRBOG: Danmarks Statistik. 1975.

WESTERGAARD.H.: Statistikkens Historie. 1931. Kap. 2.

10. BILAG.

a. Diverse konventioner og fejlkilder.

1. Personer, der har bestået socionomi, men derefter starter på et andet fag, uden at blive færdig med dette, registreres som eksamen i socionomi med studietid til eksamen. Kun i det omfang eksamen er opnået i det andet fag også, registreres det (socionomi + medie har opnået 13 stk. eksaminer). Vi har noteret de læste kombinationer senere i bilaget.

2. Bemærk, at en person kan have opnået studietrin 3 ved ét gymnasiefag, både når 2 og når 3 moduleksaminer er taget. Se senere i bilaget.

3. Nye uddannelser (f.eks. kombinationsuddannelser) er ikke kommet med. Generelt har vi pr. 1/9 1984 ikke noteret OB-fag for nytilmeldte på OB (alle øvrige har selvfølgelig OB-fag med).

4. Tilgangen til socionomi blev lukket pr. 1/9 1983 og på psykologi pr. 1/9 1982.

5. Personer, der har studeret 1 måned eller derunder er ikke regnet med i undersøgelsen.

6. Personer, der har læst på RUC (i mere end 1 måned), men udmelder sig, for senere at vende tilbage får noteret den udmeldte periode som orlov og starttidspunktet bibeholdes.

7. Personer, der tager orlov og derefter melder sig ud, er søgt meldt ud ved orlovs begyndelse.

8. Der opstår let fejl ved studietider og orlovsperioder, fordi en orlov i efterårssemestret er af 5 mdr.s længde og en orlov i foråret er på 7 måneder. Når datoerne er omregnet fra dag, måned til kun måned har vi søgt at runde op/ned; mens nogle noteringer

i de håndførte registre udelukkende rundede op og andre kun ned.

Ved udmeldelser p.gr.a. årskortets manglende fornyelse er datoerne spredt lidt på august og september af samme grund.

9. Jobmodulet (pædagogikum) er ikke regnet med i studietiderne; men der kan være opstået fejl, således at nogle har fået noteret disse 6 måneder ind i studietiden.

10. Da vi afsluttede vores register med 1/9 1984 påførte vi alle aktive deres opnåede studietrin, hvilket ellers kun var blevet påført ved afgang uden eksamen.

11. Alle er noteret med start i sept. det pågældende år, selvom nogle få starter i februar. For folk med start i januar 1973 gælder dog, at de er noteret med start i 1972.

(i) Socionomi med andre fag.

Vi har i hånden noteret 26 personer, der med bestået socionomi søgte at læse videre med et yderligere fag, men som det ikke lykkedes at fuldføre (pr. 1/9 1984). De fordelte sig således:

Andet fag:	Antal studerende	Gnst. studietid på det andet fag i måneder:
Teksam	2	30
To gymnasiefag	2	30
Forvaltning	10	16,5
Medie	12	30,5

(ii) Studietrin.

Ved tilskrivning af studietrin har vi modereret i Danmarks Statistiks metode. Vil tilskriver ikke studietrin i basis - modsat DS. En student, der har bestået basis, får studietrin 1 - hvis denne starter på overbygningen gives der 2.

Studietrinene er således for teksam, socionomi og forvaltning:

	Start	Bestået modul 1	Bestået modul 2	b.m.3	b.m.4	b.m.5
Teksam.	2	2	3	4	5	6
Forvaltning	2	2	3	4	5	6
Socionomi	2	2	3			

Med andre ord kan man have studietrin 2 både når man har bestået modul 1 og når man lige er startet på uddannelsen.

For gymnasielærerfagene kan der kun ske fejlvurdering, når man har afsluttet et af fagene ($).

	Start	Bestået modul 1	Bestået modul 2	Bestået modul 3
Fag 1	1	2	3	3($)
Fag 2	1	2	3	3($)

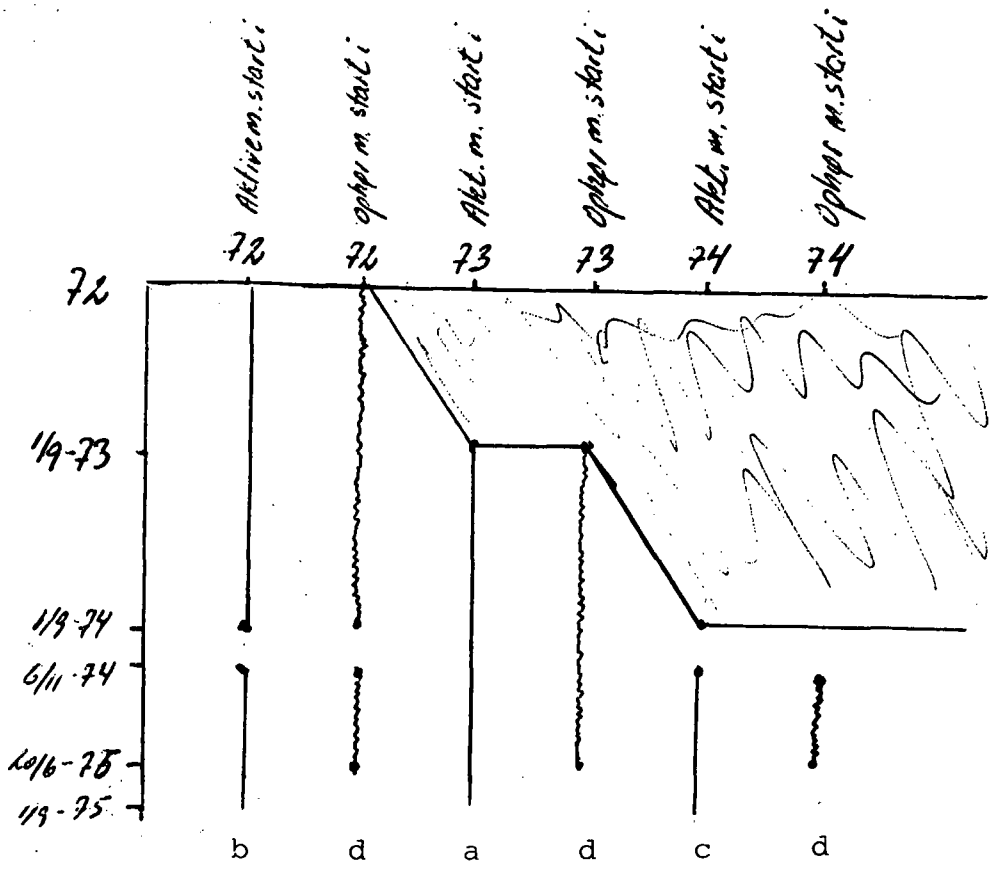
De to fags studietrin er i vores materiale blot lagt sammen. Ved trin 6 kan man altså mangle 2 moduler, eller bare mangle 1 uden at vi kan skelne.

(iii) Opstarten med matrikeludskrifterne.

For at komme igang og få dækket alle ind fra de første tre år (der blev dengang ikke lavet en systematisk matrikeludskrift hvert år) brugte vi tre matrikeludskrifter:

1. Nyimmatrikulerede ml. 1/9-73 og 1/10-73 (udskrevet d. 10/10-75)
2. Samtlige aktuelle i CPR-orden (udskrevet d. 6/11-74)
3. Samtlige med ophør på RUC

Figur 10.3 Illustration af dataindsamling fra matriklen.



På figuren herover kan det ses hvordan vi med vores procedure har fanget alle studerende fra 72 - 75.
Udskrift 1 gav os streg nr. a; alle der startede i 73
Udskrift 2 gav os yderligere streg nr. b og c; alle aktuelle d. 6/11 74
Udskrift 3 gav os stregerne d; alle udmeldte indtil 20/6 75.

På grundlag af 1) fik vi indtastet samtlige startende i 73. Fra 2) fik vi samtlige aktuelle på RUC den 6/11-74, dvs. at vi fandt ud af, hvem af 73´erne, der er gået ud mellem 1/10-73 og 6/11-74 og fandt alle øvrige aktive fra 72 og 74; men samtidig fik vi ikke dem ind fra 72 og 74, som var gået ud inden den 6/11-74. For at få fat i "ophørerne" brugte vi 3) som desværre kun har den oplysning, at et CPR-nr er gået ud og ikke hvilket år dette CPR-nr startede. Det betyder, at dem, som stoppede mellem 1/9-74 og 6/11-74 kan have startet på RUC enten i 72 eller 74.

Men disse personer har vi noteret og fået dem "undersøgt" på dataterminalen på matrikelkontoret. De personer, som holdt op inden 1/9-73 må naturligvis være 72'ere, mens dem mellem 1/9-73 og 1/9-74 var en 72'er eller 73'er; men da vi havde samtlige 73'ere inde, kunne vi nemt skelne mellem dem. Det største problem med dem, som startede i 72 og holdt op inden 20/6-75, var, at vi ikke vidste hvilken basis-uddannelse de havde haft, så her tildelte vi dem en basis efter det forhold, som udgående inden basiseksamen for de tre basisretninger havde: Hum 123 stk., Nat 47 stk., Sam 72 stk. (1), hvilket omtrent svarer til forholdet 5:2:3; som vi så tillagde de udgåede studerende.

b. Div. tilladelser og brevvekslingen desangående

For at kunne bruge matrikeloplysningerne og indberetningerne var det nødvendigt at få tilladelse fra registertilsynet. Grunden til at vi ikke umiddelbart kunne gå i gang, lå i, at det er personoplysninger vi arbejdede med og at vi jo oprettede vores eget "lille" register over RUC´s studerende.

Proceduren for at få tilladelse til at benytte sådanne oplysninger er at vi søgte RUC om tilladelse til at benytte RUC´s oplysninger. Hvis RUC giver tilladelse så sender universitetet ansøgningen + RUC´s accept videre til registertilsynet, som har den endelige afgørelse.

For at få tilladelsen blev vi rådet til at beskrive vores arbejde som forskning, hvad der jo ikke er løgn, samt at lægge vægt på, at man ikke i resultatet kan analysere sig frem til enkeltpersoner - men det har også hele tiden været vores klare mål. Vores ansøgning (se nedenfor) blev sendt afsted, sammen med Jørgen Larsens anbefaling (se nedenfor) til rektor, som sendte dem videre sammen med deres egen anbefaling (se nedenfor). Efter ca. 1½ måned fik vi tilladelse

(1): Oplysningerne kommer fra RUC's studiestatistik 1982.

til at benytte oplysningerne under visse forbehold (se nedenfor) som vi dog sagtens kunne honorere.

IMFUFA ROSKILDE UNIVERSITETSCENTER
INSTITUT FOR STUDIET AF MATEMATIK OG FYSIK SAMT DERES FUNKTIONER I UNDERVISNING, FORSKNING OG ANVENDELSER

til rektor, RUC.

POSTBOX 260
DK-4000 ROSKILDE
DANMARK

TELEFON 02-757711
LOKAL:

23.11.1984
DATO/REFERENCE JOURNALNUMMER DERES REFERENCE/JOURNALNUMMER

Ansøgning om tilladelse til at benytte RUC's matrikel-oplysninger siden starten i 1972 til nu til brug for forskningsprojekt på matematik, RUC.

Vi er tre matematikstuderende, der hermed vil ansøge om tilladelse til at benytte RUC's studenteroplysninger. Vi vil i vort afsluttende matematikprojekt studere statistiske modeller i forskningsøjemed. Ved hjælp af den såkaldte Cox-regressionsmodel håber vi at kunne afdække RUC-studenternes studiemønstre, og derigennem også teste den statistiske models anvendelse på denne type data.

Modellen er gennem de senere år blevet anvendt indenfor lægevidenskabelig forskning, hvor man har testet forskellige behandlinger af patienter. Ofte har man brugt tiden indtil dødsfald som den ønskede teststørrelse og derfor er denne type modeller hyppigt blevet kaldt overlevelsesmodeller.

Imidlertid vil oplysninger om RUC-studerendes studietider (regnet frem til eksamen eller udmeldelse) være helt oplagte for denne model. Til det formål vil vi få brug for persondata på hver enkelt studerendes 1) køn 2) alder 3) basisuddannelse 4) overbygningsuddannelser 5) orlovsperioder og 6) tidspunkt for eksamen/udmeldelse.

Disse oplysninger vil efter EDB-behandling i den stati-

stiske model kunne kortlægge studiemønstre på RUC i en generaliseret form. Vi vil kunne undersøge: Studerer kvinder længere end mænd, er studietiden på matematik kortere end på historie, spiller basis-valg en rolle for studietiden og har orlov en uheldig indflydelse på om man får eksamen eller ej.

Som det fremgår vil ingen persondata kunne identificeres ud fra vores resultater.

Vores forskning kan muligvis anvendes af RUC-administrationen til planlægningsformål.

Med venlig hilsen

Poul Kattler Mikael Johansen Torben Andreasen

hus 17.2, RUC.

IMFUFA ROSKILDE UNIVERSITETSCENTER
INSTITUT FOR STUDIET AF MATEMATIK OG FYSIK SAMT DERES FUNKTIONER I UNDERVISNING, FORSKNING OG ANVENDELSER

POSTBOX 260
DK-4000 ROSKILDE
DANMARK

TELEFON 02-757711
LOKAL:

til rektor, RUC.

DATO/REFERENCE | JOURNALNUMMER | DERES REFERENCE/JOURNALNUMMER

23.11.84

Vedr. ansøgning om tilladelse til at benytte RUC's matrikel-oplysninger.

Som vejleder for de studerende skal jeg hermed anbefale, at de studerende får lov til at benytte de omtalte oplysninger til deres projektarbejde.

Det vil være af stor undervisningsmæssig betydning og derudover i al almindelighed særdeles interessant at få mulighed for at analysere et datamateriale som det omtalte ved hjælp af de mest moderne matematisk-statistiske metoder. Så vidt vides er der ikke tidligere foretaget tilsvarende analyser og den her planlagte undersøgelse er derfor et originalt forskningsprojekt.

Med venlig hilsen

Jørgen Larsen

Jørgen Larsen.

ROSKILDE UNIVERSITETSCENTER
MARBJERGVEJ 35

AKADEMISK FORVALTNING
POSTBOX 260 DK-4000 ROSKILDE TELEFON 02 - 75 77 11

Registertilsynet
Gyldenløvesgade 19, 4.
1600 København V

DATO/REFERENCE	JOURNALNUMMER	DERES REFERENCE/JOURNALNUMMER
28. november 1984 JP/tj	00-130/1	

./. Roskilde Universitetscenter videresender hermed ansøgning fra 3 matematikstuderende om tilladelse til at benytte studenteroplysninger omfattet af registerforskrifter nr. 23 af 26.8.80 og nr. 16 af 8.12.81 for matrikel og eksamensregister for Roskilde Universitetscenter til brug ved et konkret projekt omfattende statistiske modellers anvendelighed på denne type data.

Da centret i forskellige sammenhænge udarbejder lignende statistikker som de i ansøgningen anførte til brug ved administrative og planlægningsmæssige formål, er man stærkt interesseret i at få en vurdering af den anvendte metodes egnethed.

Praktisk vil projektet kunne gennemføres ved, at de studerende får adgang til at benytte studenteroplysninger fra udskrifter af registret. Tilsynet med udskrifternes anvendelse varetages i det daglige af matriklens personale under opsyn af den registeransvarlige myndighed.

P.R.V.
E.B.

Jette Petersen

cc:Boel Jørgensen
INFO
EE
Matr.
Jørgen Larsen

ROSKILDE UNIVERSITETSCENTER
MARBJERGVEJ 35

POSTBOX 260 DK-4000 ROSKILDE TELEFON 02 - 75 77 11

Poul Kattler
Mikael Johansen
Torben Andreasen

Hus 17.2 gr. 5.

DATO/REFERENCE	JOURNALNUMMER	DERES REFERENCE/JOURNALNUMMER
4/2 1985	00-130/1	

I anledning af at Registertilsynet ved skrivelse af 29 januar 1985 har givet tilladelse til, at RUC videregiver oplysninger fra matrikel- og eksamensregister til jer til brug ved et konkret forskningsprojekt, anmodes I om omgående at kontakte undertegnede. Jeg vil da på et møde redegøre for de vilkår for videregivelsen, der er stillet af Registertilsynet.

Med venlig hilsen
Bette Petersen

cc: Jørgen Larsen
Matr.
JP
EE

c. En ekskurs ud i den geometriske fortolkning.

Vi står overfor to sæt af problemløsninger, der hver har sin geometrikse fortolkning. Dels Newton-Raphson iterationen, der maksimaliserer log likelihood funktionen (afhængig af $\underline{\beta}$) og dels eliminations/ og selektions-proceduren, der søger at beskrive virkeligheden med færrest mulige parametre. Den første er gentaget mange gange inde i den anden, og den anden gentages også det nødvendige antal gange, før vi er færdige med modellen.

Newton-Raphson iterationen arbejder i et n+1 dimensionelt rum, når der er n parametre. Den sidste dimension udtrykker log likelihood værdien. Vi arbejder med l' (første afledte af log likelihood funktionen), der sættes lig 0, for dermed at kunne maksimalisere de n β-værdiers billedpunkt på den n+1´te akse. Den anden afledte l'' er i denne sammenhæng altid negativ, dvs. med negativ krumning i alle dimensioner. Derfor kan vi opfatte log likelihood funktionen som et n+1 dimensionelt bjerg med netop ét toppunkt i "højden" udtrykt med log likelihood værdien. Bjerget er pænt krummet i alle retninger.

Ved hjælp af et sæt startværdier (der angiver et startpunkt) vandrer iterationen afsted mod toppen som en anden bjergbestiger i den stejleste retning. Under et vist antal trin skal både højden og placeringen af toppunktet (beskrevet af de n β-akser) konvergere. Hvis toppunktet er tilstrækkeligt fladt i en eller flere retninger kan der godt ske det, at højden findes (med en fastlagt sikkerhed); uden at det nøjagtige sæt koordinater (β-værdier) kan findes indenfor de på forhånd fastsatte trin og halveringer.

De estimerede beta-værdier vil være det bedste skøn over toppens placering i de "kendte" dimensioner.

Det udelukker ikke, at det virkelige toppunkt er et andet sted, højere oppe, beskrevet af endnu flere retninger/parametre.

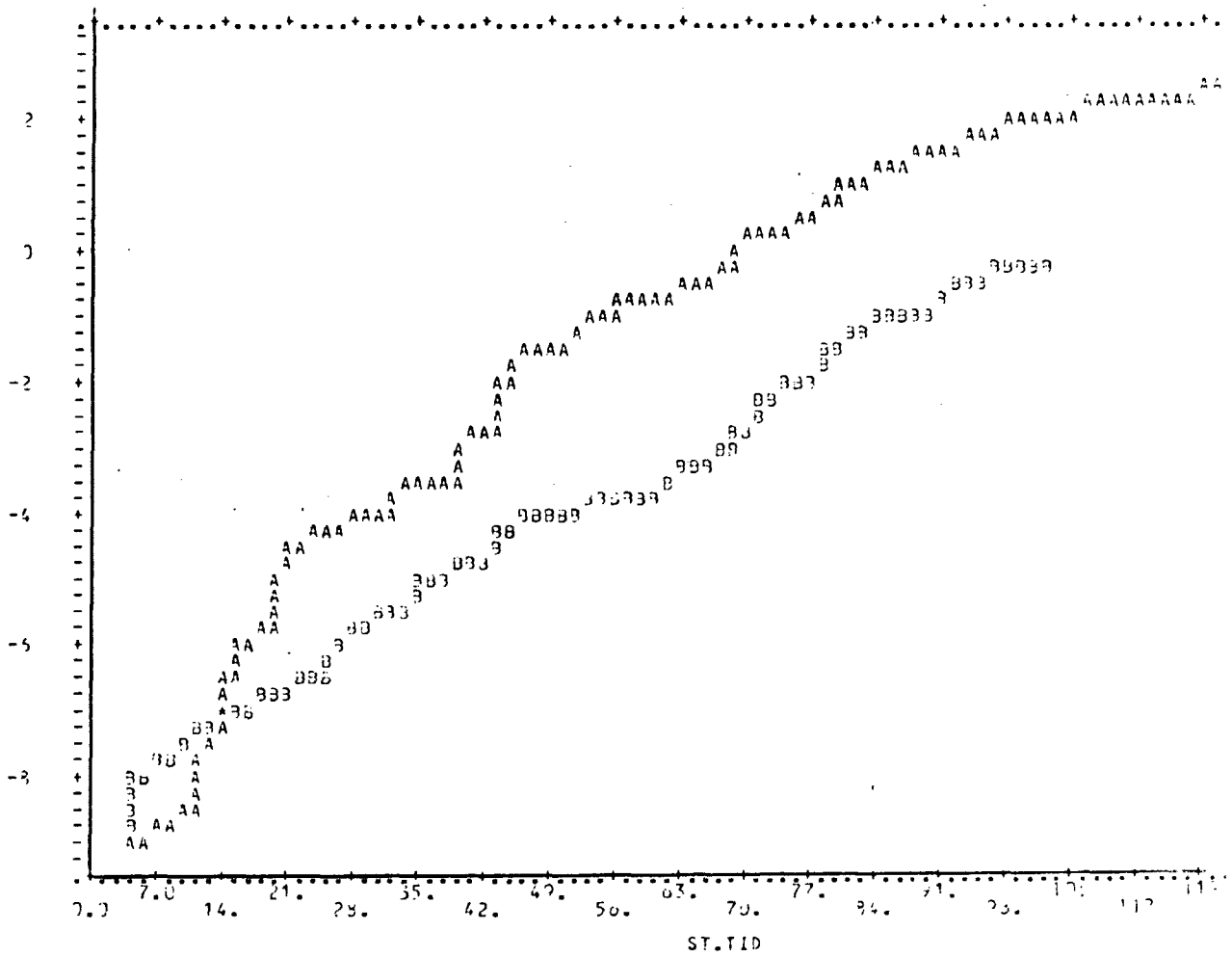
Eliminations/ og selektions-proceduren kan beskrives i et n-dimensionelt vektorrum, der udspændes af de n forklarende parametre. Parameter-vektorernes længde fastsættes numerisk af Newton-Raphson iterationen. De er alle lineært uafhængige, idet vi fra starten af har sikret os, at ingen retninger (to og to) er korreleret med faktoren +1 eller -1. De udspænder et delrum af virkeligheden, men hvis vi skal gøre en model ud af det, håber vi, at de dækker de væsentligste retninger og at vi til og med kan beskrive virkeligheds-rummet med nogle færre end de n til rådighed værende parametre. Afhængig af forward eller backward procedurer udvælger vi det nødvendige og tilstrækkelige antal parametre. Ved forward proceduren lægger vi de givne vektorer ind - først dén der ligger i den mest fremherskende retning i virkeligheds-figuren, så dén der peger i den næst-vigtigste retning, osv. Vi prøver at indfange rumfanget af en mange-dimensionel "cigar", hvilket gøres bedst og simplest ved hjælp af de mest vinkelrette retninger. Da vores vektorer imidlertid ikke er vinkelrette (de er korrelerede) vil vi være tvunget til at vælge flere end "nødvendigt", idet den bedste eller simpleste beskrivelse ville kunne gøres v.hj.a. de mest ukorrelerede parametre (vinkelrette vektorer). Når vi finder frem til k væsentlige parametre (af længderne β_1 til β_k), ville vi uden tvivl med andre og bedre parametre kunne finde en forklaring med et mindre antal parametre, der i så fald var tilstrækkeligt mindre korrelerede. Imidlertid har vi kun rådighed over de n kendte parametre, samt deres vekselvirkninger (der kan udvide vektorrummet vilkårligt, men i ret "kedelige" retninger). Ved backward elimination ser vi på de arealer, vektorerne indbyrdes to og to udspænder og fjerner skridt

for skridt den mindste vektor fra det par, der udspænder det mindste areal.

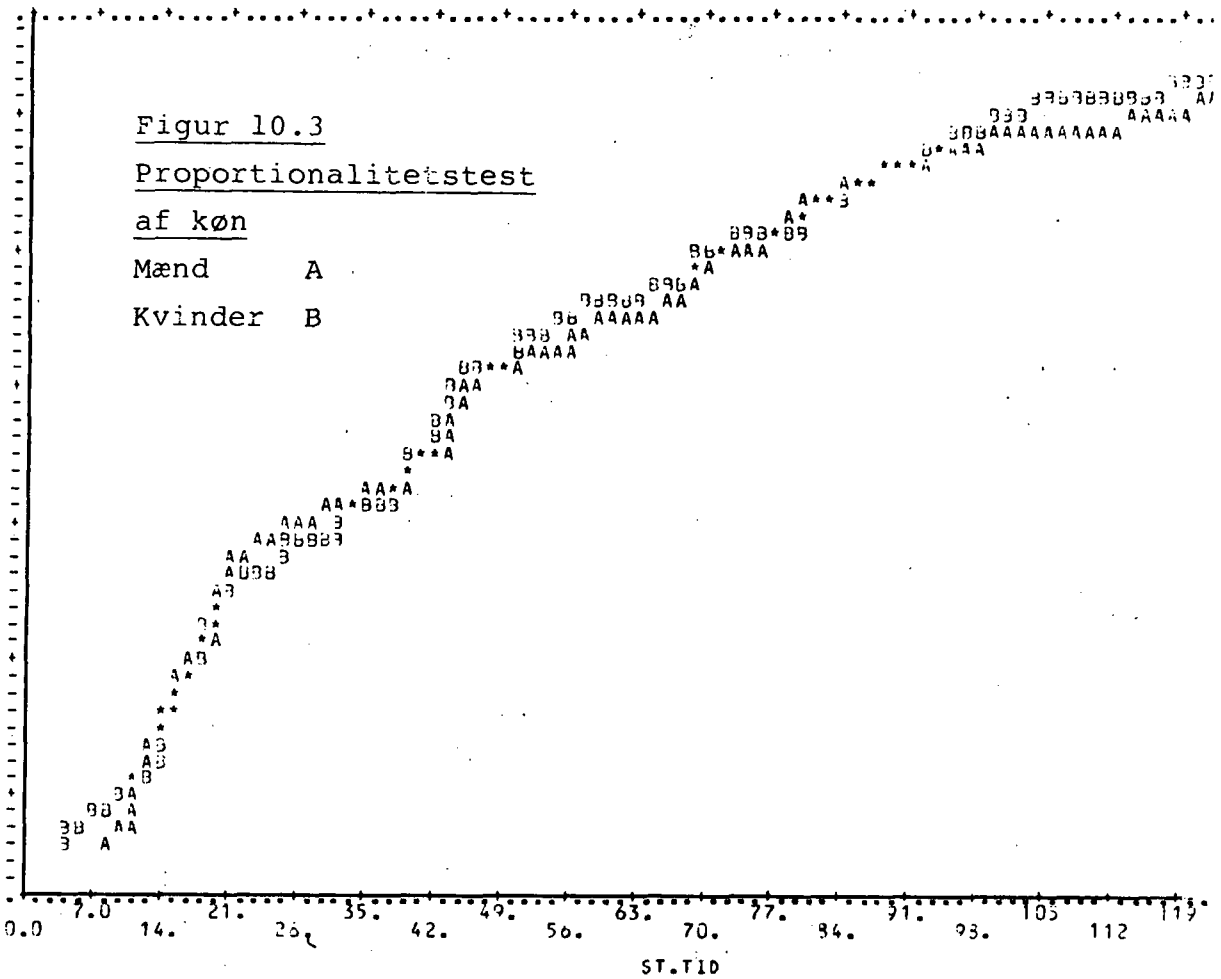
Imellem hvert skridt (uafhængig af metoden) må vi på ny udregne korrelationer og beta-værdier for at bestemme det bedste næste skridt. Vi stopper proceduren, når forbedringen/forværringen overskrider nogle på forhånd fastsatte værdier.

d. Figurer fra kapitel 7, der ikke er placeret i teksten.

Figur 10.2 Proportionalitetstest af forvaltning (B) og øvrige fag (A).



Figur 10.3
Proportionalitetstest
af køn
Mænd A
Kvinder B



Figur 10.4
Proportionalitetstest
af alder