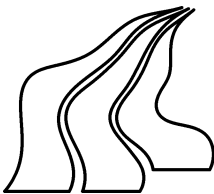


Fed fisks betydning for livvidde-ændringer

- modelleret ved hjælp af blandede modeller og metaanalyse

Speciale projekt
Foråret 2010
Karen Margrete Due



INSTITUT FOR MATEMATISKE FAG, AALBORG UNIVERSITET
KARDIOVASKULÆRT FORSKNINGSCENTER, AALBORG SYGEHUS,
ÅRHUS UNIVERSITETS HOSPITAL

Titel:

Fed fisks betydning for
livvidde-ændringer
- modelleret ved hjælp af
blandede modeller og
metaanalyse

Projektperiode:

SUMA10,
1. feb. 2010 - 1. jun. 2010

Forfatter:

Karen Margrete Due

Vejleder:

Claus Dethlefsen

Oplagstal: 6

Sidetal: 141

Afsluttet: 1. jun. 2010

*Offentliggørelse af rapporten må
kun ske efter aftale med forfat-
teren.*

Synopsis:

I dette projekt undersøges sammenhængen mellem livvidde-ændringer, i forhold til højde og indtag af fed fisk. Til at undersøge denne sammenhæng, benyttes data fra The European Prospective Investigation into Cancer and Nutrition studie.

Det kan diskuteres, hvorvidt der er tale om ét stort studie eller 27 mindre studier. Derfor fokuserer dette projekt på forskellige statistiske metoder til, at modellere ét samlet estimat for den ønskede analyse på trods af, at data ikke nødvendigvis stammer fra det samme studie. Der betragtes to metoder: Blandede modeller samt metaanalyse. Metoderne sammenlignes og deres styrker og svagheder diskuteres.

Da det ikke a priori forventes, at sammenhængen mellem indtag af fed fisk og ændringer i livvidde, i forhold til højde, er lineær, betragtes ligeledes to metoder til ikke-lineær modellering af kontinuerte variable. Disse metoder er restringerede kubiske splines og fraktionelle polynomier. Metoderne inkorporeres ligeledes i metaanalyserne.

Projektet er inddelt i tre del. Første del omhandler den egentlige undersøgelse af livvidde-ændringer i forhold til højde og indtag af fed fisk, anden del gennemgår den basale teori anvendt til analyserne, medens tredje del består af en beskrivelse af de udviklede statistiske programmer. Programmerne anvendes igennem projektet til grafiske illustrationer.

FORORD

Denne rapport er udarbejdet som et sundhedsmatematisk speciale hos Institut for Matematiske Fag, ved Aalborg Universitet, i perioden 1. februar til 1. juni 2010. Projektet er udarbejdet i samarbejde med Kardiovaskulært Forskningscenter, ved Aalborg sygehus - Århus universitets hospital. Dele af kapitel 2, 3 og 10 tager udgangspunkt i delafsnit udarbejdet i løbet af 9. semester, i perioden fra 1. september til 21. december 2009.

En liste over notation anvendt i dette speciale findes i appendix A.

Litteraturhenvisninger er noteret som $[x]$, hvor x er et tal som henviser til litteraturlisten, som findes sidst i rapporten.

Aalborg, 1. juni 2010.

Karen Margrete Due

INDHOLD

1	Summary in English	5
2	Introduktion	9
I		13
3	Datapræsentation	15
3.1	Sammenligning af studier	16
3.1.1	Studiepopulation	17
3.1.2	Kost oplysninger	17
3.1.3	Livsstilsfaktorer og antropometri	17
3.2	Valg af studie population	19
4	Opstilling af modeller	25
4.1	Troværdighedsmål	26
4.2	Analysestrategi	27
4.3	Optimering af modellerne	30
4.3.1	Valg af antal knudepunkter	30
4.3.2	Valg af potens og grad	33
4.4	Valg af model	37
5	Resultater	45
5.1	Resultater - Kvinder	45
5.2	Resultater - Mænd	53
6	Konklusion	57
II		61
7	Lineære blandede modeller	63
7.1	Modeller med fast effekt	63
7.2	Modeller med tilfældig effekt	65

7.2.1	Troværdighedsmålet	68
7.3	Blandede modeller	69
8	Parameter estimation	71
8.1	Maksimum likelihood estimation	71
8.1.1	Akaike's informations kriterium	74
8.2	Restringeret maksimum likelihood	75
9	Metaanalyse	79
9.1	Univariat metaanalyse med fast effekt	79
9.2	Univariat metaanalyse med tilfældig effekt	83
9.2.1	Estimation af de ukendte parametre	84
9.3	Multivariat metaanalyse med fast effekt	88
9.4	Multivariat metaanalyse med tilfældig effekt	91
9.4.1	Estimation af ukendte parametre	91
9.5	Perfekt prædiktion	94
9.6	Metaanalyse i forhold til blandede modeller	94
10	Ikke-lineær modellering af kontinuerte variabler	97
10.1	Restringerede kubiske splines	99
10.1.1	Valg af antallet af knudepunkter	101
10.1.2	Bestemmelse af konfidensbånd	102
10.2	Fraktionelle polynomier	107
10.2.1	Valg af potens samt grad af det fraktionelle poly- nomium	110
10.2.2	Bestemmelse af konfidensbånd	114
10.3	Diskussion	116
A	Notation	121
B	Centerspecifikke resultater - kvinder	123
C	Centerspecifikke resultater - mænd	124
D	Hjælpefil for <code>splot_make</code>	125
E	Hjælpefil for <code>splot</code>	129
F	Hjælpefil for <code>mvfrac</code>	132

G Hjælpefil for mvspline	135
Litteratur	138

SUMMARY IN ENGLISH

The aim of this thesis is to investigate the link between intake of fatty fish and the change in waist circumference in relation to height. To examine this relationship we use data from The European Prospective Investigation into Cancer and Nutrition study collected from 27 research centers in ten European countries in the period from 1993 to 1999. To investigate the hypothesis of this project we have construct some exclusion criteria's to modify the data for our thesis. After the modification there are 74,650 observations left, of which 57 percent are women and 43 percent are men.

The thesis consists of three parts. The first part deals with the use of mixed effects models and meta analysis to investigate the hypothesis of this project and description of the used dataset. The hypothesis of this project is that the intake of fatty fish has an reducing effect on abdominal obesity. Were abdominal obesity is expressed as the relationship between changes in waist circumference and height. The hypothesis concerning the beneficial effect of consuming fish on abdominal obesity is established because other studies have shown an indication of, that fish protein may have an effect on satiety [18] [34] and the subsequent energy intake [2], compared to other protein sources. Furthermore fish contains the ω -3 fatty acids eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), which are polyunsaturated fatty acids. According to [22] the development of obesity in rodents is held back by adding ω -3 fatty acids to their diet.

Furthermore the strategy of analysis is outlined in the first part of the report and appropriate models are selected. The two different analytical strategies, mixed effects models and meta analysis, are used because it is debatable whether the data can be perceived as one large study or a number of smaller studies. Since we do not a priori expect the relationship between intake of fatty fish and changes in Waist circumference in relation to height, to be linear, we use non-linear modelling for the intake of fatty fish as our starting point for modelling

the effect of intake of fatty fish. For this purpose we used two different methods: restricted cubic splines and fractional polynomials. The results of the two methods are compared and eventual differences are pointed out. All models are stratified by sex because we in previous cross-sectional studies have experienced that the intake of fish affects the waist circumference of women and men differently. We therefore expected the same in the follow up study. The analysis for both sexes shows no significant effect of increasing the intake of fatty fish. The exposure variable fatty fish were modelled as a fractional polynomial for women and as linear for men. The models have been chosen by using deviance difference, deviance tests and Akaike's information criteria. The expression of the models is found in chapter 4 expression (4.1) page 38 and (4.2) page 41. The results of the model for females (4.1) are illustrated graphically because the coefficients for non-linear modelling can not be interpreted separately. The graph is found in Figure 5.2 page 48. The results for the model for men (4.2) are presented as the regression coefficient for increasing the intake of fatty fish by 10 grams per day. The results are shown in Table 5.1. Even though the results are not significant the tendency for both sexes is that by increasing the intake of fatty fish, the increasing of waist circumference in relation to height is reduced.

The second part of the thesis consists of the basic theory behind mixed effect models, meta analysis, parameter estimation, restricted cubic splines and fractional polynomials. The methods mixed effect models and meta analysis is reviewed and their advantages and disadvantages are discussed. We prefer mixed effect models for the analysis of our problem, but believe that meta analysis is an effective tool in regards to assessing whether many studies can be considered as a comprehensive study. Similarly the methods restricted cubic splines and fractional polynomials are reviewed and their advantages and disadvantages are discussed. Our evaluation is that it depends on the situation which of the two methods are preferable, but that restricted cubic splines are preferable to remove the confounding effects while fractional polynomials is suitable for modelling of exposure variables where the curves are controlled.

The third part of the thesis is not a self-sufficient part of the report. During the period of the project we have developed three additional programs intended for STATA 11, which is the statistical program we have used for our analysis. One of the programs is a program for drawing restricted cubic spline plots. The program is called *splot* and was produced because there are non built-in function in STATA 11 which can easily draw spline plots. Besides, we have written two programs for combining meta analysis with restricted cubic splines and fractional polynomials. These types of programs do not currently exist to our knowledge. The coefficients from restricted cubic splines and fractional polynomials can not be interpreted individually. To interpret the results the functions have to be presented as a graph. The features *mv spline* and *mvfrac* create this type of graphs, respectively for each study individually and for the weighted average found by using meta analysis. All programs that have been written in the context of this thesis can be found at <http://kmd.kvfc.dk/>.

INTRODUKTION

Kraftig overvægt er et stort problem i vores samfund. Problemet er så stort, at der ifølge WHO [37] er tale om en europæisk fedmeepidemi. Fedme er en væsentlig underliggende risikofaktor for en række sygdomme, blandt andet hjerte- og kredsløbssygdomme, alders diabetes [32] og kræft [4].

Intet tyder på, at antallet af kraftigt overvægtige, vil blive mindre i fremtiden. Det anslås, at hvis den nuværende stigende tendens ikke vendes, kan fedme blive den mest alvorlige sundhedstrussel i den vestlige verden, kun overgået af tobaksrygning. Forebyggelse af fedme er derfor vigtigt set udfra et folkesundhedsmæssigt synspunkt.

Fedme er forårsaget af ubalance mellem energiindtag og energiforbrug [27]. Fedme bliver dog også påvirket af andre livsstilsfaktorer, såsom rygning og alkohol. Hvilke faktorer der påvirker overvægt, kan ligeledes variere. Eksempelvis for forskellige europæiske områder, befolkningsgrupper, mellem mænd og kvinder samt i aldersgrupper.

En større forståelse af årsagsmekanismer for fedme, i forskellige europæiske befolkninger, vil skabe et solidt videnskabeligt grundlag for, at udvikle mere specifikke retningslinjer, til at forebygge og mindske fedmetendenser i hele Europa. [10]

I 1993 blev The European Prospective Investigation into Cancer and Nutrition (EPIC) opstartet. EPIC er et stort internationalt multi-center-samarbejde etableret til, at studere kroniske sygdomme, som har stor betydning for folkesundheden. Det vil sige sygdomme såsom kræft, hjerte-karsygdomme og diabetes. EPIC er den største undersøgelse af kost, biomarkører og sundhed, der hidtil er blevet gennemført. Der er rekrutteret over en halv million mennesker fra 27 centre i ti europæiske lande. [10]

Med udgangspunkt i data fra EPIC databasen, vil vi forsøge, at belyse

hvordan fedme påvirkes af indtag af fed fisk, hvor fedme udtrykkes ved forholdet mellem livvidde og højde (livvidde(mm)/højde(m)). Dette mål er ikke et ofte anvendt mål for fedme. Målet anvendes, idet abdominalfedme anses for at være mere skadeligt end andre typer af fedme. Fedtaflejring omkring maven er tættere associeret med flere kroniske sygdomme end eksempelvis fedt aflejret på hofterne [16]. Dette kan skyldes, at fedtceller i bughulen er metabolisk mere aktive end fedtceller placeret andre steder [3]. Ved metabolisme i fedtvævet udskilles triglycerider. Når metabolismen i fedtcellerne i bughulen er mere aktive, udskilles altså flere triglycerider [1]. Øget antal triglycerider kan blandt andet medføre koronararterie sygdomme, slagtilfælde og diabetes [16]. Ifølge [20] giver forholdet mellem livvidde og højde en mere præcis beskrivelse af fordelingen af fedt på kroppen end body mass indexet $\left(\text{vægt(kg)} / (\text{højde(m)})^2 \right)$, idet BMI kun tager højde for vægt og ikke hvor på kroppen fedtet er placeret. Flere studier her desuden vist, at forholdet mellem livvidde og højde er bedre til at forudsige kardiovaskulære risikofaktorer samt død, end andre antropometriske mål for fedme [6], [17] og [19]. Derfor betragtes i dette projekt markøren ændringer i livvidde/højde.

Hypotesen for dette projekt er, at indtag af fed fisk har en reducerende effekt på abdominalfedme. Hvor abdominalfedme er udtrykt ved forholdet mellem ændringer i livvidde i forhold til højde. Til at understøtte denne hypotese, betragtes andre relaterede studier.

I studierne [18] og [34] betragtes følelsen af mæthed efter indtag af fiskeproteiner, fremfor andre proteinkilder. I [34] sammenlignes indtag af fisk med indtag af oksekød og kylling, mens fiskeindtaget i [18] sammenlignes med 38 forskellige madvarer. I begge studier betragtes deltagernes egen fornemmelse af mæthed, som viser sig at være signifikant større efter indtag af fisk. I [34] er referencegrupperne oksekød og kylling, mens referencegruppen i [18] er hvidt brød. Et tredje studie, som har betragtet sammenhængen mellem fiskeprotein og mæthed er [2]. Dette studie betragter fiskeprotein i forhold til protein fra oksekød. Til forskel fra de to foregående studier, betragter [2] udover mæthed ligeledes det efterfølgende energiindtag. Studiet [2] fandt en tendens til at mætheden efter fiskeindtag er større, men ikke

signifikant. Imidlertid fandt de at gennemsnits energi indtaget efter indtag af fisk var 2.805 ± 1.095 kilojoule sammenlignet med 3.133 ± 904 kilojoule efter indtag af oksekød. Hvilket giver et signifikant lavere energiindtag efter indtag af fisk.

Den direkte sammenhæng mellem fedme og indtag af ω -3 fedtsyrer, som er de fedtsyrer der er i fisk, er blevet undersøgt hos gnavere. I [22] betragtes tilførslen af ω -3 fedtsyrer i gnaveres kost. Ifølge [22] er ω -3 fedtsyrer bedre til at aktivere PPAR- α receptorerne end andre fedtsyrer, hvilket medfører en reduceret fedtmasse hos gnavere. PPAR- α er en del af PPAR-familien, som er en gruppe af nukleare receptor proteiner, bestående af PPAR- α , PPAR- β/δ og PPAR- γ . Alle tre PPAR'er er involverede i regulering af opbygning og nedbrydning af lipider, som er en samlet betegnelse for fedtstof. Hver af PPAR'ene har forskellige roller. PPAR- γ findes i fedtceller samt tyktarmen og indgår i dannelsen af adipocytter, som er de celler, der er specialiseret i lagring af energi som fedt, samt i adipocyt differentiationen, som er deling af adipocyt cellerne. PPAR- β/δ findes i næsten alle væv og indgår i de indledende stadier af adipocyt differentiationen. En forøgelse af PPAR- β/δ og PPAR- γ vil intuitivt give en forøget fedtmasse. PPAR- α findes i lever, nyrer, hjerte og muskler og indgår i modsætning til PPAR- β/δ og PPAR- γ i fedtsyrer nedbrydningen. Studierne betragtet i [22] viste en forøgelse af PPAR- α samt en reduktion af PPAR- γ ved tilførsel af ω -3 fedtsyrer.

Mængden af ω -3 fedtsyrer i fisk afgør, hvorvidt der er tale om en mager eller en fed fisk. Er der mere end otte gram ω -3 per 100 gram fisk, er der ifølge [12] tale om fede fisk. Disse er for eksempel makrel, laks, sild, ansjoser og sardiner. Da indholdet af ω -3 fedtsyrer er størst i fede fisk, betragtes i dette projekt fede fisks indvirkning på ændringer i forholdet mellem livvidde og højde.

Da EPIC undersøgelserne stammer fra 27 centre, kan det diskuteres hvorvidt der er tale om et stort studie eller 27 mindre studier. Til at undersøge projektets hypotese, anvendes derfor to metoder: metaanalyse og blandede modeller. Hvor metaanalyse betragter hvert af de 27 centre som separate studier og udfra de enkelte estimater

finder et samlet estimat for sammenhængen mellem indtag af fed fisk og ændringer i livvidde/højde, mens blandede modeller opstiller en samlet model, som tager højde for centerforskellene. Formålet med dette projekt er, udover at undersøge hypotesen vedrørende indtag af fed fisk og ændringer i livvidde i forhold til højde, at studere de to analyse metoder samt sammenligne deres styrker og svagheder.

Derudover betragtes i dette projekt to metoder til ikke-lineær modellering af kontinuerte variabler. Disse er : restringerede kubiske splines og fraktionelle polynomier. De to metoder inkorporeres ligeledes i metaanalyserne, hvilket efter vores kendskab, ikke er muligt på nuværende tidspunkt i de statistiske programmer der er til rådighed. Ikke-lineær modellering betragtes, idet vi ikke a priori forventer at sammenhængen mellem indtag af fed fisk og ændringer i livvidde/højde er lineær.

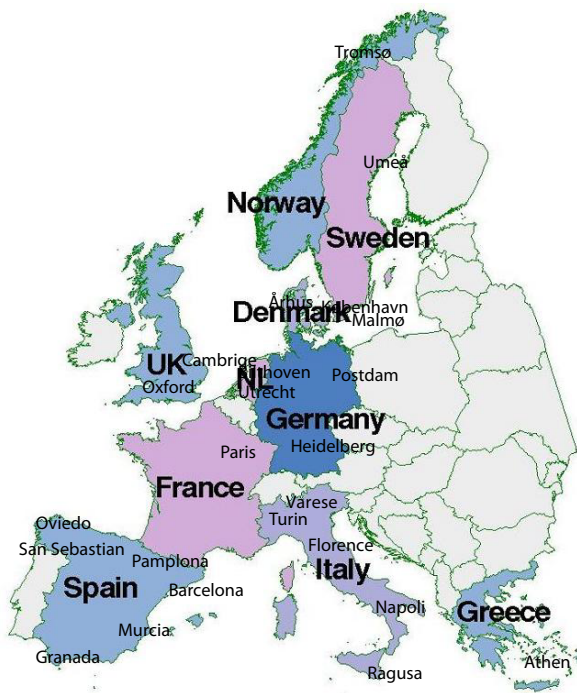
Projektet er opdelt i tre dele. Første del indeholder den praktiske del, herunder en introduktion til data fra EPIC, gennemgang af kendte risikofaktorer for fedme, valg af studiepopulation og en redegørelse for valg af modeller samt analyser og resultater for disse. I anden del redegøres for teorien bag analysemetoderne, som er anvendt i første del. Tredje del består af udvikling af nødvendige programmer til STATA 11, som er det statistiske program, som anvendes til analyser i dette projekt.

Del I

DATAPRÆSENTATION

The European Prospective Investigation into Cancer and Nutrition (EPIC), er et prospektivt kohorte studie, som har til formål at undersøge forholdet mellem kost og kræft, såvel som andre kroniske sygdomme, såsom hjerte-kar-sygdomme. Med over en halv million deltagere, er EPIC på verdensplan blandt de største undersøgelser af kost og sygdom, der på nuværende tidspunkt, er iværksat. EPIC er finansieret af Europe Against Cancer programmet, ledet af Europa kommissionen, samt en række nationers specifikke tilskud og velgørende organisationer. [10]

Informationer vedrørende deltagerne er indsamlet i 27 centre, fordelt over ti europæiske lande. Hvilke lande samt placeringen af centrene, ses i Figur 3.1. Rekruttering til undersøgelsen fandt sted mellem 1993 og 1999. Opfølgningen var planlagt til at vare mindst ti år. Nogle af deltagerne er blevet fulgt i de planlagte ti år, mens andre er faldet fra undervejs. Opfølgningen er sket via registerdatabaser og interview/spørgeskemaer, hvert tredje til femte år. Der er indsamlet data i form af standardiserede kostspørgeskemaer, syv-dages mad dagbøger, 24-timers recall, livsstilsspørgeskemaer, blodprøver, antropometriske målinger, såsom body mass index og talje-til-hofte ratio samt register oplysninger hvor det var muligt. Hvilket alene for kostfrekvensspørgeskemaerne og livsstilsspørgeskemaer svarer til cirka 1000 oplysninger per deltager, hvortil register oplysninger og oplysninger fra biologisk materiale skal tilføjes.



Figur 3.1: I figuren ses hvilke lande og centre der er med i EPIC.

3.1 Sammenligning af studier

Alle de 27 studier er opstartet med samme formål, nemlig at undersøge forholdet mellem kost og kræft. På trods af at studierne har samme formål, kan det diskuteres, hvorvidt man kan betragte de 27 studier som et stort studie eller om de skal betragtes som 27 mindre studier. Denne problemstilling har udspring i at:

- I Studiepopulationerne ikke er ens i alle studierne.
- II Spørgsmålene vedrørende kost ikke er helt ens i alle landene. Hvilket blandt andet skyldes at kosten varierer fra land til land.
- III Ikke alle centre har de samme oplysninger vedrørende livsstilsfaktorer.

3.1.1 Studiepopulation

I Tabel 3.1 er de vigtigste forskelle mellem studiepopulationerne listet op. For en mere detaljeret beskrivelse af de 27 centre se [29]. I tabellen ses det, at de fleste af studiepopulationerne er indsamlet som populations-baserede studier, i større eller mindre grad. Studiepopulationerne er som oftest begrænset af bopælsområde og alder. At studierne er populations-baserede betyder i denne sammenhæng, at en del af populationen er blevet inviteret og de som ønskede at deltage er med i studiet. Vi forventer ikke at der opstår selektionsbias i denne forbindelse, idet vi regner med at indtag af fed fisk har samme påvirkning uafhængigt af, om man har valgt at deltage i studiet eller ej. Visse af studierne skiller sig ud ved, at de eksempelvis kun har kvinder eller vegetarer med. Alderen af deltagerne ved studiestart er for de fleste studier mellem 35 og 65 år, men fire studier har individer ned til 20 år med.

3.1.2 Kost oplysninger

Oplysningerne vedrørende kost er primært indsamlet ved hjælp af kostfrekvensspørgeskemaer. Flere forskellige udformninger af kostfrekvensspørgeskemaer er blev brugt i EPIC, blandt andet fordi kosten varierer fra land til land. For at gøre kostoplysningerne sammenlignelige er der blevet udarbejdet en fælles standard kategorisering i kostgrupper, på tværs af landene i EPIC. Visse oplysninger er på trods af standardiseringen ikke lige detaljerede i alle studierne. Blandt andet er fisk ikke opdelt i mager og fed fisk for studierne Potsdam (Tyskland), Heidelberg (Tyskland), Murcia (Spanien) og Napales (Italien).

3.1.3 Livsstilsfaktorer og antropometri

De fleste af centrene har de samme oplysninger om antropometri enten som selv-rapporteret eller målte, men for centrene Umeå (Sverige) og Tromsø (Norge) er der kun information om vægt og højde. Der kan være stor forskel på de antropometriske målinger, alt efter hvordan de er målt. Nogle er målt og vejjet med tøj medens andre

3. DATAPRÆSENTATION

Land	Center	Studiepopulation	Alder
Grækenland	Athen	Populations-baseret	25-82 år
	Spanien	Bloddonorer og populations-baseret	35-64 år
	Granada	Bloddonorer og deres partnere,	35-65 år
	Murcia	populations-baseret, embedsmænd, ansatte ved to firmaer samt deltager fra et andet studie	
	Pamplona	Bloddonorer og populations-baseret	35-65 år
	San Sebastian	Bloddonorer og ansatte fra udvalgte firmaer	35-65 år
	Ovido	Bloddonorer, embedsmænd og populations-baseret	35-64 år
	Italien	Bloddonorer, populations-baseret, medlemmer af lærer foreningen samt andre kilder	35-65 år
	Ragusa	Personer der er blevet scannet for brystkræft og populations-baseret	35-64 år
	Florence	Bloddonorer, ansatte, frivillige samt testpersoner af medicin	35-74 år
	Turin	Frivillige	35-65 år
	Varese	Kvindelige frivillige	30-69 år
Frankrig	Napals	National sundhedsforsikring, lærer og andre ansatte på skoler	40-65 år
	Paris	Populations-baseret	35-65 år
Tyskland	Heidelberg	Populations-baseret	20-65 år
	Postsam	Populations-baseret	20-65 år
	Holland	Populations-baseret	20-65 år
	Amsterdam	Populations-baseret	20-65 år
	Doetinchem	Populations-baseret	49-70 år
	Maastricht	Personer der er blevet scannet for brystkræft	45-74 år
	Utrecht	Patienter	(1) 40-65 år
England	Cambrigde/Norfolk	(2) vegetarer, veganere og andre helbredsbevidste personer	(2) 20+ år
	Oxford	Populations-baseret	50-64 år
Danmark	Århus	Populations-baseret	50-64 år
Sverige	København	Populations-baseret	46-72 år
	Malmø	Populations-baseret	30,40,
	Umeå	Populations-baseret	50 eller 60 år
	Norge	Populations-baseret	født mellem 1943 og 1957

Tabel 3.1: I tabellen ses, hvordan studiepopulationerne i de forskellige studier er udvalgt samt hvordan aldersfordelingen er indenfor hvert af disse.

uden, i nogle centre er højden blevet målt både stående og siddende men ikke i andre, derudover kommer tillige, at nogle deltagere selv har målt og vejset sig medens andre er blevet målt. Tendensen til at de fleste centre har sammenlignelige oplysninger, men at få skiller sig ud, går igen indenfor flere områder i EPIC. Blandt andet kender man kun information om nuværende alkoholindtag for deltagere fra Malmø (Sverige) og Tromsø (Norge), hvor man for de andre centre ligeledes kender tidligere alkohol indtag. Informationer om fysisk aktivitet er for Umeå (Sverige) og Tromsø (Norge) er ikke sammenlignelige, med oplysningerne om fysisk aktivitet fra de andre studier.

Grunden til at man ønsker at betragte de mange studier som ét

studie, på trods af deres forskelle er, at det giver en højre statistisk styrke, hvormed man kan estimere effekter, af eksempelvis indtag af fed fisk, med større præcision. For at man kan betragte studierne som et stor studie, bør de være sammenlignelige. Vi vil derfor udvælge en studiepopulation vi finder sammenlignelige til, at betragte vores problemstilling.

3.2 Valg af studie population

Den fulde EPIC kohorte består af 521.448 individer. Til at betragte problemstillingen vedrørende livsstilsfaktorerens betydning for ændringer i livvidde i forhold til højde, opstilles en række eksklusionskriterier. Disse kriterier har til formål at udvælge de studier som kan betragtes som sammenlignelige i forhold til problemstillingen samt sikre at der er data til rådighed for de involverede variabler. Eksklusionskriterierne tager udgangspunkt i eksklusionskriterierne fra Diet, Obesity, and Genes (DiOGenes)[8]. DiOGenes består af en række undersøgelser, som ligeledes tager udgangspunkt i data fra EPIC. De første ni af de efterfølgende eksklusionskriterier er overført fra DiOGenes, mens de sidste to er specifikke for dette projekt:

- I Livviddemålinger ved opfølgnings tidspunktet mangler.
- II Længden af opfølgningstiden er nul.
- III Graviditet.
- IV Intet kostspørgeskema.
- V Ekstrem ratio mellem rapporteret energiindtag og forventet energiforbrug.
- VI Ingen anthropometri mål ved studiestart, herunder livvidde og højde.
- VII Ingen opfølgningstid eller antropometriske mål ved opfølgningstidspunktet.

VIII Upålidelige anthrometri ved både studiestart og opfølgningstidspunktet: højde $<1,3\text{m}$, BMI $<16\text{ kg/m}^2$, livvidde $<40\text{ cm}$ eller livvidde $>160\text{ cm}$ eller livvidde $<60\text{ cm}$ mens BMI $>25\text{ kg/m}^2$.

IX Upålidelige ændringer i vægt eller højde: $< -5\text{ kg/år}$ eller $> +5\text{ kg/år}$, eller $< -7\text{ cm/år}$ eller $> 7\text{ cm/år}$.

X Mangler information om potentielle confoundere.

XI Har ikke oplysninger om fisk opdelt i mager og fed fisk.

De potentielle confoundere for sammenhængen mellem indtag af fed fisk og ændringer i livvidde i forhold til højde kan opdeles i tre kategorier: Kendte risikofaktorer, demografiske variabler og forklaringsvariabler.

Kendte risikofaktorer for fedme

- Socialstatus: Lav socialstatus er associeret med fedme [33]. Socialstatus udtrykkes her ved uddannelsesniveau.
- Rygning/rygeændringer: Rygning er negativt associeret med fedme. Ved rygestop øges risikoen for fedme [23].
- Fysisk aktivitet: Risikoen for fedme reduceres ved øget fysisk aktivitet [28].
- Menopausalstatus: Når kvinder kommer i overgangsalderen, har de tendens til forøget vægt [24].

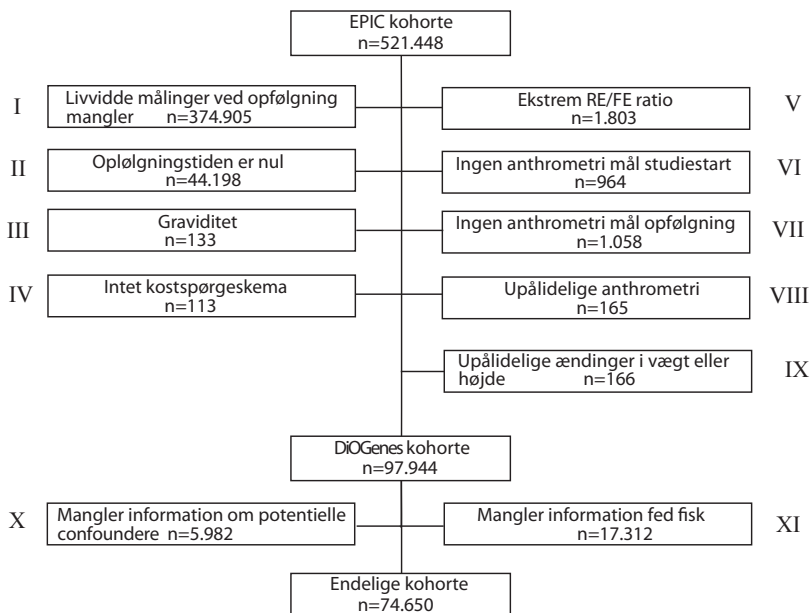
Demografiske variabler

- Center
- Land
- Alder
- Køn

Forklaringsvariabler

- **Opfølgningsperiode:** Det er forskel på hvor lang deltagernes opfølgningsperiode er. For at de livvidde-ændringer der betragtes er sket indenfor lige lang tid, er det nødvendigt, at justere for opfølgningstid.
- **Energi:** Ved at justere for energi sammenlignes deltagere som indtager samme mængde energi. En forøgelse af indtaget af fed fisk sker derved på bekostning af andre fødevarer.

Hvor mange deltagere der fjernes på grund af eksklussionskriterierne, ses flowchartet i Figur 3.2. Eksklusionerne sker i samme rækkefølge, som de er listet i de ovenstående eksklussionskriterierne **I-XI**.



Figur 3.2: I figuren ses, hvor mange deltagere, der ekskluderes på baggrund af de enkelte eksklussionskriterier samt, hvor mange der er tilbage i den endelige kohorte.

3. DATAPRÆSENTATION

Efter alle eksekutionerne er der kun information fra syv centre tilbage: 1. Firenze (Italien) 2. Cambridge (England) 3. Amsterdam (Holland) 4. Doetinchem (Holland) 5. Maastricht (Holland) 6. Århus (Danmark) 7. København (Danmark). Fordelingen af data for den tilbageværende populationen, Tabel 3.2. For de tilbageværende studier er livviddemålingerne selv-rapporteret for Firenze medens de er blevet målt for de andre centre.

Center

n

74.650

manglende

0

unikke

7

Firenze

9.585

13

Cambridge

12.658

17

Amsterdam

2.553

3

Doetinchem

3.558

5

Maastricht

2.904

4

Århus

13.344

18

København

30.048

40

Frekvens

%

Køn

n

74.650

manglende

0

unikke

2

Mænd (32.213, 43%), Kvinder (42.437, 57%)

Ændring i livvidde i forhold til højde (mm/m)

Ryge status

n	manglende	unikke
74.650	0	4

Ikke-ryger (52.476, 70%), Ryger (15.615, 21%), Begyndt at ryge (1.370, 2%)
Stoppet med at ryge (5.189, 7%)

Fysisk aktivitet

n	manglende	unikke
74.650	0	4

Inaktive (11.461, 15%), Moderat inaktive (23.551, 32%)
Moderat aktive (17.650, 24%), Aktive (21.988, 29%)

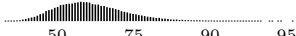
Menopausalstatus

n	manglende	unikke
42.437	32.213	4

Premenopausal (8.556, 20%), Postmenopausal (25.035, 59%)
Perimenopausal (7.037, 17%)
Kirurgisk postmenopausal (1.809, 4%)

Energi indtag

n	manglende	unikke	gns.	.05	.10	.25	.50	.75	.90	.95
74.650	0	74.394	2.177	1.330	1.472	1.750	2.109	2.531	2.966	3.258



Tabel 3.2: I tabellen ses fordelingen af de potentielle confoundere, eksponeringsvariablen fed fisk og responsvariablen ændringer i livvidde i forhold til højde.

Fordelingen af den observerede ændring i livvidde/højde i forhold til indtag af fed fisk ses for hvert center i Figur 3.3. Det ses, at du fra det observerede forhold mellem livvidde-ændringer i forhold til højde og indtag af fed fisk, ser det umiddelbart ikke ud til, at der er en sammenhæng. Derudover ses det, at Cambridge ikke har samme detaljeringsgrad som de andre studier, hvorfor der vil blive taget særlig hensyn til Cambridge i modelopstillingsfasen.

Desuden ses det, at deltagerne i de danske studier rapporterer et større fiskeindtag end deltagerne i de andre studier. Dette skyldes muligvis, at der er flere spørgsmål vedrørende fisk i det danske kostfrekvensspørgeskema. For at rette op på denne type målefejl, er der udarbejdet forskellige kostkalibreringsmetoder. Disse vil ikke blive betragtet i dette projekt, men en beskrivelse findes i [9]. Når der ikke justeres for målefejl, kan der argumenteres for at resultaterne i virkeligheden afspejler ændringer i livvidde/højde i forhold til hvad deltagerne rapporterer de spiser, fremfor hvad de i virkeligheden spiser.

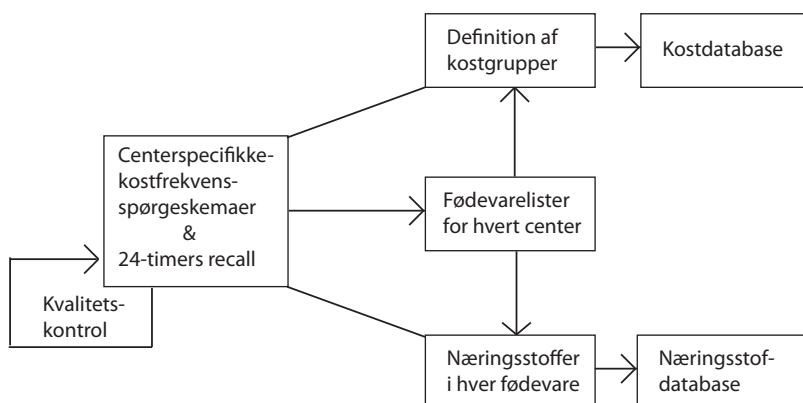
OPSTILLING AF MODELLER

Vi er i dette projekt interesseret i, at betragte sammenhængen mellem responsvariablen ændringer i livvidde/højde og den forklarende variabel, indtag af fed fisk, med udgangspunkt i datasættet fra EPIC, hvor eksklusionerne er foretaget. Til at modellere sammenhængen, anvendes metoderne blandede modeller, som beskrives i kapitel 7 og metaanalyse, som beskrives i kapitel 9. Begge metoder er statistiske metoder, som anvendes til, at modellere et samlet estimat for forholdet mellem responsvariablen og eksponeringsvariablen, når data er inddelt i flere studier.

Responsvariablen er i dette tilfælde ændringer i livvidde i forhold til højde, hvor livvidden er målt i millimeter og højden er målt i meter. Da højden ikke ændrer sig i løbet af opfølgningsperioden, er højden målt ved studiestart og livvidden er målt ved studiestart og opfølgningstidspunktet. Ændringen i livvidde i forhold til højde er da fundet ved:

$$\Delta \frac{\text{Livvidde (mm)}}{\text{Højde(m)}} = \frac{\text{Livvidde ved opfølgningstid(mm)}}{\text{Højde ved studiestart(m)}} - \frac{\text{Livvidde ved studiestart(mm)}}{\text{Højde ved studiestart(m)}}$$

Eksponeringsvariablen er indtag af fed fisk i gram per dag. Indtag af fed fisk er målt ved hjælp af kostfrekvensspørgeskemaer som efterfølgende er blevet bearbejdet i et fælles program til at lave næringsstofberegninger. Programmet kan omregne de retter og specifikke fødevarergrupper deltagerne har opgivet, til mere overordnede grupper så som kød, mejeriprodukter, grønsager, frugt og så videre og helt ned til ernæringsniveau. Hvordan man kommer fra kostfrekvensspørgeskemaerne til kostgrupper og næringsstoffer er skematiseret i Figur 4.1. I Tabel 3.2 så vi at størstedelen af deltagernes indtag af fed fisk lå mellem 0 og 50 gram per dag. Spiser man en halv rugbrød med fiskepålæg svarer det cirka til at indtage 25 gram fed fisk, se [11].



Figur 4.1: I figuren ses en skematisering af hvordan man kommer fra de enkelte kostfrekvensspørgeskemaer til et samlet estimat for kostgrupper eller næringsstoffer.

4.1 Troværdighedsmål

Inden vi går i gang med at opstille modeller for problemstillingen, udregnes troværdighedsmålet, jævnfør afsnit 7.2.1, for at sikre, at de modeller der opstilles, ikke bliver for komplekse i forhold til den datainformation vi har til rådighed. Da indtag af fed fisk er en kontinuert variabel, anvendes troværdighedsmålet $p < n/20$. Troværdighedsmålet udregnes separat for mænd og kvinder, idet vi laver kønsstratificerede analyser. Troværdighedsmålet for henholdsvis kvinder og mænd, ses i Tabel 4.1. Hvis antallet af fri parametre i modellen er mindre end $n/20$ er modellen troværdig i forhold til overfitting, ifølge troværdighedsmålet. [14]

	n	$n/20$
Kvinder	42.437	2.122
Mænd	32.213	1.611

Tabel 4.1: I tabellen ses troværdighedsmålet for henholdsvis kvinder og mænd.

4.2 Analysestrategi

Modellerne opstilles med udgangspunkt i de potentielle confoundere, beskrevet i kapitel 3. For at se om det har en effekt, at justere for de forklarende variabler, laves en ujusteret analyse som sammenlignes med de tilsvarende justerede analyser.

For begge analysemetoder gælder ligeledes, at analyserne tager højde for grupperingerne i centre. I de blandede modeller medtages centrene som en tilfældig effekt, mens der i metaanalyserne er stratificeret på center. I alle modellerne anvendes de samme potentielle confoundere. Udfra definitionen af de to metoder, svarer fortolkningen af justeringen i metaanalyserne til, at der indføres interaktion mellem alle forklarende variabler og centervariablen i de tilsvarende blandede modeller.

Alle analyser laves separat for kvinder og mænd, idet vi i tværsnitsstudierne [9] så en forskel i, hvordan mænd og kvinders livvidde blev påvirket af indtag af fisk. Vi forventer derfor, at effekten af indtag af fed fisk på ændringer i livvidde/højde ligeledes er forskellig for mænd og kvinder. Vi har valgt at stratificere fremfor at lave en interaktion med køn, idet tendensen for sammenhængen mellem livvidde og fisk i [9] var modsatrettede.

De resterende potentielle confoundere: Socialstatus, rygeændringer, fysisk aktivitet, menopausalstatus for kvinderne, alder, opfølgningsperiode og energi, medtages for at reducere eventuel confounding effekt. Derfor justeres der for variablerne så godt som data tillader. Variablerne Socialstatus, rygeændringer, fysisk aktivitet og menopausalstatus er kategoriske variabler og justeres derfor for som en række indikatorvariabler svarende til antallet af kategorier. Antallet af kategorier for hver af variablerne kan ses i Tabel 3.2. Første kategori sættes til referenceniveau og de resterende indekseres med $\mathbb{I}_{\text{variabel},j}$ for den j 'te kategori. De resterende forklarende variabler, alder, opfølgningsperiode og energi, er kontinuerte variabler og vil blive justeret for som restringerede kubiske splines med syv knudepunkter. Hvis der ifølge troværdighedsmålet er risiko for overfitting nedjusteres antallet af knudepunkter. For de restringerede kubiske splines genereres en variabel som indeholder formen af tredjegradspolynomiet

mellem to knudepunkter. Variablen defineres som det ses i udtryk (10.1) side 99 og variablen for det j 'te interval indekseres $x_{\text{variabel}'j}$. Vi betragter i dette studie både restringerede kubiske splines og fraktionelle polynomier til ikke-lineær modellering af kontinuerte variabler. Vi har valgt at medtage de kontinuerte forklarende variabler som restringerede kubiske splines, idet restringerede kubiske splines ligger tættere op ad data end fraktionelle polynomier og derved fjerner mere af confounder effekten. For uddybning af metoderne restringerede kubiske splines og fraktionelle polynomier samt forskellene mellem disse, se kapitel 10.

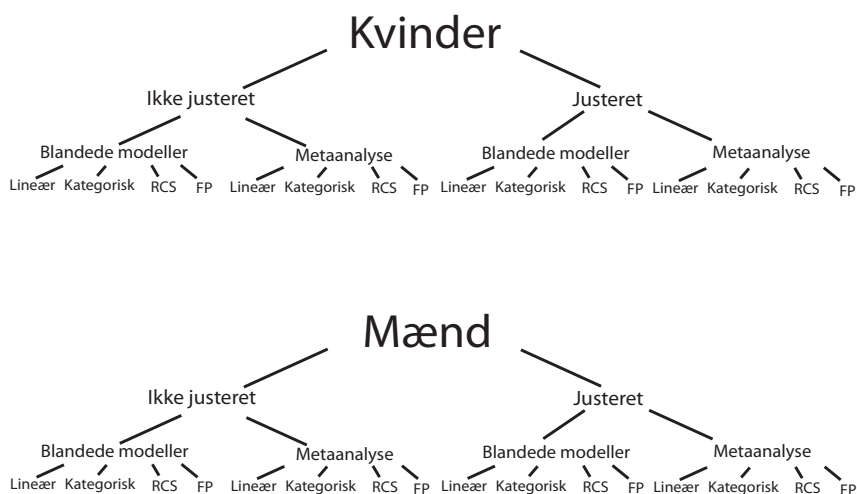
Eksponeringsvariablen indtag af fed fisk modelleres som henholdsvis lineær, inddelt i kvintiler og som ikke-lineær henholdsvis som restringerede kubiske splines og fraktionelle polynomier. Modellerne hvor fed fisk medtages som lineær og inddelt i kvintiler medtages primært for at se, hvor stor betydning det har, at vælge den rette metode til, at modellere ens variabler. I alle modellerne medtages ligeledes en indikatorvariabel som angiver om man spiser fed fisk eller ej. Denne indikatorvariabel medtages, idet vi forventer at der er forskel på at spise lidt fed fisk og slet ikke at spise fed fisk.

Ved brug af blandede modeller antages, at de grupper vi har til rådighed, er en delmængde af uendelig mange normalfordelte grupper. Vi forventer at centrene har forskellige udgangsniveauer, men at effekten af indtag af fed fisk ellers er den samme, hvorfor vi antager at interceptet for hver gruppe er trukket fra en normalfordeling. Variationen mellem studierne kan derved forklares udelukkende ud fra afvigelsen fra middelværdien. Studierne opfattes derved som en del af et større studie, hvor forskellene kan fjernes ved, at justere for oprindelses sted. Vi mener derudover, at der er en sammenhæng mellem hvilket land man kommer fra, i sammenhæng med hvilket center man hører til. Derfor modellerer vi sammenhængen mellem ændringer i livvidde/højde og indtag af fed fisk, med en blandet model for indlejrede data, hvor center er indlejrede i land. Modellering af indlejret data beskrives i afsnit 7.2.

Vi fortrækker blandede modeller, da de kan give mere fleksible

modellerings muligheder. Vi mener ligeledes at man med fordel kan betragte metaanalyserne, blandt andet til at vurdere, hvorvidt man kan betragte centrene som en samlet population eller som flere mindre populationer. Vi laver derfor alle analyserne både ved hjælp af blandede modeller og metaanalyse. Resultaterne fra de blandede modeller er hovedresultaterne, mens metaanalyserne sammen med test for heterogenitet, i det univariate tilfælde, anvendes til, at vurdere hvorvidt det er fornuftigt at lave et samlet estimat.

Hvis alle modeller laves, ender vi ud med i alt 32 modeller, som fordeler sig som det ses i Figur 4.2. I de ikke-justerede modeller indgår kun eksponeringsvariablen modelleres som enten lineær, kategorisk, en restringeret kubisk spline eller et fraktionelt polynomium, samt en indikatorvariabel for indtag af fed fisk eller ej. For indikatorvariablen sættes referenceniveauet til, at man spiser fisk. For de justerede modeller justeres ligeledes for alle de potentielle confoundere.



Figur 4.2: I figuren ses, hvordan de 32 potentielle modeller fordeler sig.

4.3 Optimering af modellerne

Der findes ingen korrekte modeller. Man kan kun finde den der passer bedst i forhold til de betingelser man stiller op. Når man vælger sin model, er det derfor en balance mellem præcision i henhold til det enkelte datasæt og generaliserbarhed.

Vi har valgt at anvende de tre mål, devians differensen, devians test og Akaike's informations kriterium til, at bestemme den optimale model. For målene devians differensen og Akaike's informations kriterium gælder, at jo højere værdi, der fås des bedre er modellen. Ved devians testet anvendes signifikansniveau $\alpha = 0,05$.

For at vælge de mest optimale modeller bestemmes først, ved hjælp af devians differensen, hvilket antal knudepunkter der er optimal for modellen, hvor eksponeringsvariablen modelleres som en restringeret kubisk spline. Dernæst bestemmes, ligeledes ved hjælp af devians differens, hvad der er den optimale potens, for modellen hvor eksponeringsvariablen modelleres som et fraktionelt polynomium. Sidst bestemmes den optimale grad for det fraktionelle polynomium ved hjælp af devians test. Når antallet af knudepunkter samt grad og potens er bestemt, anvendes Akaike's informations kriterium til, at vælge imellem de lineære, kategoriske og ikke-lineære modeller. Optimeringen udføres for de blandede modeller for såvel for de ikke-justerede som de justerede modeller.

4.3.1 Valg af antal knudepunkter

I Tabel 4.2 ses devians differensen, udregnet mellem den lineære model og modellen hvor maksimum likelihood estimatet er anvendt, for hver af modellerne, hvor eksponeringsvariablen er modelleret som en restringeret kubisk spline, med $k = 3, \dots, 7$. Devians differensen anvendes til at vælge hvor mange knudepunkter modellen, for henholdsvis mænd og kvinder, skal have. Det ses, at for både den justerede og ikke-justerede model er den mest optimale model, for kvinder, modellen hvor syv knudepunkter anvendes. For mændene er den model, som får den højeste devians differens modellen hvor

seks knudepunkter anvendes. Vi har derfor valgt at anvende syv knudepunkter for modellerne for kvinderne og seks for modellerne for mændene.

	Kvinder		Mænd	
	ikke-justeret	Justeret	ikke-justeret	Justeret
Knudepunkter	Dev. diff.	Dev. diff.	Dev. diff.	Dev. diff.
k=3	-5,2	4,6	0,2	0,2
k=4	-5	4,6	0,4	0,4
k=5	-5	4,6	0,6	0,2
k=6	-4,4	5,8	5,4	4,4
k=7	-2,8	7,6	5,2	3,8

Tabel 4.2: I tabellen ses devians differensen for alle modellerne, hvor eksponeringsvariabel fed fisk er modelleret ved hjælp af restringerede kubiske splines.

Som det beskrives i afsnit 10.1, kan antallet af knudepunkter teoretisk øges uden grænse, men [14] anbefaler man anvender tre til syv knudepunkter. Da vi for kvinderne fik den største devians differens i grænsetilfældet, hvor antallet af knudepunkter er syv, kan man overveje om en model med endnu flere knudepunkter ville give en endnu højere devians differens. Da troværdighedsmålet for kvinder tillader at der anvendes 2.122 parametre i modellen, vil de ekstra knudepunkter og dermed ekstra parametre, der skal estimeres, ikke give problemer i forhold til overfitting. Grunden til vi på trods af dette har valgt, at stoppe ved syv knudepunkter er, at de redskaber der er til rådighed i STATA 11, kun kan finde restringerede kubiske splines med tre til syv knudepunkter. En anden grund til ikke at øge antallet af knudepunkter er, at jo flere knudepunkter der anvendes, des mere datanært modelleres eksponeringsvariablen og dermed fås en model, som passer meget godt til det datasæt man betragter, men måske ikke er så generaliserbart.

Knudepunkternes placering er bestemt som i Tabel 10.1 side 101. Da indtag af fed fisk ligeledes modelleres ved hjælp af en indikatorvariabel,

er knudepunkterne fundet for alle med et positivt indtag af fed fisk. Derudover er Cambridge taget ud ved bestemmelse af knudepunkterne, idet Cambridge udgør 17 procent af den samlede datamængde, men kun har syv mulige udfald, hvorfor Cambridge vil præge placeringen af knudepunkterne mere end ønsket. Knudepunkternes placering ses for henholdsvis kvinder og mænd i Tabel 4.3.

Køn	Indtag af fed fisk (g/dag)						
Kvinder (k=7)	0,5	3,6	7,2	11,1	15,7	22,6	23,5
Mænd (k=6)	1,0	5,9	11,5	17,2	25,9	27,4	

Tabel 4.3: I tabellen ses, hvor knudepunkterne, anvendt til modellering af de restringerede kubiske splines, er placeret for henholdsvis kvinder og mænd.

Forskriften for fed fisk modelleret som restringerede kubiske splines bliver da for en kvinde:

$$\begin{aligned}
 f(x) = & \beta_0 + \beta_1 x_{\text{fed fisk}} \\
 & + \beta_2 \left((x_{\text{fed fisk}} - 0,5)^3 - (x_{\text{fed fisk}} - 22,6)^3 + \frac{25,5 - 0,5}{23,5 - 22,6} + (x_{\text{fed fisk}} - 23,5)^3 + \frac{23 - 0,5}{23,5 - 22,6} \right) \\
 & + \beta_3 \left((x_{\text{fed fisk}} - 3,6)^3 - (x_{\text{fed fisk}} - 22,6)^3 + \frac{23,5 - 3,6}{23,5 - 22,6} + (x_{\text{fed fisk}} - 23,5)^3 + \frac{22,6 - 3,6}{23,5 - 22,6} \right) \\
 & + \beta_4 \left((x_{\text{fed fisk}} - 7,2)^3 - (x_{\text{fed fisk}} - 22,6)^3 + \frac{23,5 - 7,2}{23,5 - 22,6} + (x_{\text{fed fisk}} - 23,5)^3 + \frac{22,6 - 7,2}{23,5 - 22,6} \right) \\
 & + \beta_5 \left((x_{\text{fed fisk}} - 11,1)^3 - (x_{\text{fed fisk}} - 22,6)^3 + \frac{43 - 11,1}{23,5 - 22,6} + (x_{\text{fed fisk}} - 23,5)^3 + \frac{22,6 - 11,1}{23,5 - 22,6} \right) \\
 & + \beta_6 \left((x_{\text{fed fisk}} - 15,7)^3 - (x_{\text{fed fisk}} - 22,6)^3 + \frac{23,5 - 15,7}{23,5 - 22,6} + (x_{\text{fed fisk}} - 23,5)^3 + \frac{22,6 - 15,7}{23,5 - 22,6} \right),
 \end{aligned}$$

og for en mand:

$$\begin{aligned}
 f(x) = & \beta_0 + \beta_1 x_{\text{fed fisk}} \\
 & + \beta_2 \left((x_{\text{fed fisk}} - 1,0)^3 - (x_{\text{fed fisk}} - 25,9)^3 + \frac{27,4 - 1,0}{27,4 - 25,9} + (x_{\text{fed fisk}} - 27,4)^3 + \frac{25,9 - 1,0}{27,4 - 25,9} \right) \\
 & + \beta_3 \left((x_{\text{fed fisk}} - 5,9)^3 - (x_{\text{fed fisk}} - 25,9)^3 + \frac{27,4 - 5,9}{27,4 - 25,9} + (x_{\text{fed fisk}} - 27,4)^3 + \frac{25,9 - 5,9}{27,4 - 25,9} \right) \\
 & + \beta_4 \left((x_{\text{fed fisk}} - 11,5)^3 - (x_{\text{fed fisk}} - 25,9)^3 + \frac{27,4 - 11,5}{27,4 - 25,9} + (x_{\text{fed fisk}} - 27,4)^3 + \frac{25,9 - 11,5}{27,4 - 25,9} \right) \\
 & + \beta_5 \left((x_{\text{fed fisk}} - 17,2)^3 - (x_{\text{fed fisk}} - 25,9)^3 + \frac{27,4 - 17,2}{27,4 - 25,9} + (x_{\text{fed fisk}} - 27,4)^3 + \frac{25,9 - 17,2}{27,4 - 25,9} \right).
 \end{aligned}$$

Til modeludvælgelsen er der anvendt maksimum likelihood estimation til at estimere de ukendte parametre. Når vi har valgt vores modeller, anvendes restringeret maksimum likelihood estimation til, at estimere de ukendte parametre. Restringeret maksimum likelihood estimation foretrækkes, når varians og kovarians parametre skal estimeres, idet maksimum likelihood estimation underestimerer variansen. Grunden til vi ikke anvender restringeret maksimum likelihood estimation til modeludvælgelsen er, at ved restringeret maksimum likelihood estimation anvendes forskellige sæt af observationer for de forskellige modeller, hvormed deres middelværdistrukturer bliver forskellige og derved bliver deres log-likelihoodfunktioner ikke sammenlignelige. For uddybelse af maksimum likelihood estimation, restringeret maksimum likelihood estimation og brugen af disse, se kapitel 8.

4.3.2 Valg af potens og grad

Devians differensen anvendes ligeledes til at bestemme den mest optimale potens for modellerne, hvor eksponeringsvariabel fed fisk er modelleret ved hjælp af fraktionelle polynomier. I Tabellerne 4.4, 4.5, 4.6 og 4.7 ses devians differens mellem den lineære model og hver af de andre mulige modeller, modelleret ved hjælp af fraktionelle polynomier for henholdsvis den ikke-justerede og justerede model for kvinder og den ikke-justerede og justerede model for mænd.

Kvinder - Ikke-justeret										
FP1				FP2						
Potens q	Dev.diff	Potens		Dev.diff	Potens		Dev.diff	Potens		Dev.diff
		q_1	q_2		q_1	q_2		q_1	q_2	
-2	-0,8	-2	-2	-0,8	-1	1	0	0	2	0,6
-1	-0,8	-2	-1	0,8	-1	2	-0,6	0	3	3,6
-0,5	-0,8	-2	-0,5	3	-1	3	1,8	0,5	0,5	1,4
0	0,2	-2	0	2,4	-0,5	-0,5	3,2	0,5	1	1,8
0,5	1	-2	0,5	1	-0,5	0	2,2	0,5	2	4,2
1	0	-2	1	0	-0,5	0,5	1	0,5	3	7,8
2	-0,8	-2	2	-0,6	-0,5	1	0	1	1	2,8
3	1,8	-2	3	1,8	-0,5	2	-0,6	1	2	6
		-1	-1	1,6	-0,5	3	2	1	3	9,4
		-1	-0,5	3	0	0	1,8	2	2	9,4
		-1	0	2,4	0	0,5	1	2	3	11,8
		-1	0,5	1	0	1	0,4	3	3	12,8

Tabel 4.4: I tabellen ses devians differensen for FP1 og FP2 for den ikke-justerede model for kvinder. Den højeste devians differenserne er markeret.

4. OPSTILLING AF MODELLER

Kvinder - Justeret											
FP1				FP2							
Potens	q	Dev.diff	Potens		Dev.diff	Potens		Dev.diff	Potens		Dev.diff
			q1	q2		q1	q2		q1	q2	
	-2	0	-2	-2	0	-1	1	0	0	2	0,4
	-1	0	-2	-1	0,8	-1	2	0,4	0	3	2,4
	-0,5	0	-2	-0,5	0,8	-1	3	2,2	0,5	0,5	0
	0	0	-2	0	0,2	-0,5	-0,5	0,6	0,5	1	0
	0,5	0	-2	0,5	0	-0,5	0	0,2	0,5	2	0,8
	1	0	-2	1	0	-0,5	0,5	0	0,5	3	3,2
	2	0,4	-2	2	0,4	-0,5	1	0	1	1	0,2
	3	2,2	-2	3	2,2	-0,5	2	0,4	1	2	1,6
			-1	-1	1	-0,5	3	2,2	1	3	4,2
			-1	-0,5	0,8	0	0	0	2	2	4
			-1	0	0,2	0	0,5	0	2	3	6,6
			-1	0,5	0	0	1	0	3	3	8,4

Tabel 4.5: I tabellen ses devians differensen for FP1 og FP2 for den justerede model for kvinder. Den højeste devians differenserne er markeret.

Mænd - Ikke-justeret										
FP1		FP2								
Potens	q	Dev.diff		Potens		Dev.diff	Potens		Dev.diff	Potens
		q1	q2		q1	q2		q1	q2	
-2	-0,4	-2	-2	-0,4	-1	1	2	0	2	0,4
-1	-0,4	-2	-1	1,4	-1	2	0,8	0	3	-0,8
-0,5	-0,4	-2	-0,5	-0,2	-1	3	0	0,5	0,5	1,8
0	-1,2	-2	0	1	-0,5	-0,5	-0,4	0,5	1	0,6
0,5	-0,8	-2	0,5	2,2	-0,5	0	1	0,5	2	-0,6
1	0	-2	1	2	-0,5	0,5	2,2	0,5	3	-0,8
2	-0,6	-2	2	0,8	-0,5	1	2,2	1	1	0
3	-1,4	-2	3	0	-0,5	2	0,8	1	2	0
		-1	-1	1	-0,5	3	0	1	3	0
		-1	-0,5	-0,2	0	0	2	2	2	0
		-1	0	1	0	0,5	2,6	2	3	-0,2
		-1	0,5	2,2	0	1	2	3	3	-0,6

Tabel 4.6: I tabellen ses devians differensen for FP1 og FP2 for den ikke-justerede model for mænd. Den højeste devians differenserne er markeret.

Det ses af tabellerne, at den optimale potens for modellering af sammenhængen mellem indtag af fed fisk og ændringer i livvidde/højde er for kvinder, for den ikke-justerede model, er $q = 3$ for første grad og $\mathbf{q} = (q_1 \ q_2) = (3 \ 3)$ for anden grad.

Hvilket for en kvinde giver følgende fraktionelle polynomier:

$$f_1(x_{\text{fed fisk}}, 3) = \beta_0 + \beta_1 x_{\text{fed fisk}}^3$$

og

$$f_2(x_{\text{fed fisk}}, (3 \ 3)) = \beta_0 + \beta_1 x_{\text{fed fisk}}^3 + \beta_2 \log(x_{\text{fed fisk}}) x_{\text{fed fisk}}^3.$$

Mænd - Justeret									
FP1		FP2							
Potens q	Dev.diff	Potens		Dev.diff	Potens		Dev.diff	Potens	
		q_1	q_2		q_1	q_2		q_1	q_2
-2	0,6	-2	-2	0,6	-1	1	2,6	0	2
-1	0,6	-2	-1	2,6	-1	2	2,2	0	3
-0,5	0,6	-2	-0,5	1	-1	3	1,2	0,5	0,5
0	-0,2	-2	0	1	-0,5	-0,5	0,6	0,5	1
0,5	-1	-2	0,5	2,2	-0,5	0	1,2	0,5	2
1	0	-2	1	2,6	-0,5	0,5	2,2	0,5	3
2	0,2	-2	2	2,2	-0,5	1	2,6	1	1
3	-0,6	-2	3	1,2	-0,5	2	2,2	1	2
		-1	-1	2,2	-0,5	3	1,4	1	3
		-1	-0,5	1	0	0	2	2	2
		-1	0	1	0	0,5	2,8	2	3
		-1	0,5	2,2	0	1	3	3	3

Tabel 4.7: I tabellen ses devians differensen for FP1 og FP2 for den justerede model for mænd. Den højeste devians differenserne er markeret.

For den justerede model fås de samme optimale potenser.

For mænd fås, at $q = 1$ og $\mathbf{q} = (q_1 \ q_2) = (0 \ 0,5)$ er de optimale potenser for den ikke-justerede model.

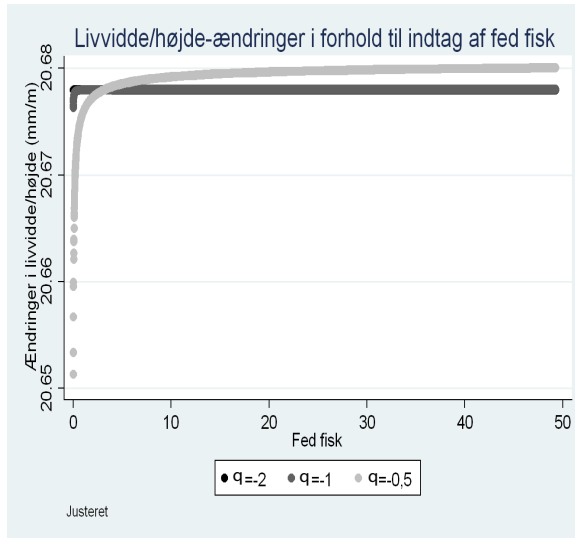
Hvilket for en mand giver følgende fraktionelle polynomier:

$$f_1(x_{\text{fed fisk}}, 1) = \beta_0 + \beta_1 x_{\text{fed fisk}}$$

og

$$f_2(x_{\text{fed fisk}}, (0 \ 0,5)) = \beta_0 + \beta_1 \log(x_{\text{fed fisk}}) + \sqrt{x_{\text{fed fisk}}}.$$

For den justerede model fås, for første grad, at $q = -2$, $q = -1$ og $q = -0,5$ giver samme devians differens for den justerede model mens $\mathbf{q} = (q_1 \ q_2) = (0 \ 1)$ er optimal for det fraktionelle polynomium af anden grad. Vi har indtegnet de tre modeller for det fraktionelle polynomium af første grad for $q = -2$, $q = -1$ og $q = -0,5$ i Figur 4.3. Det ses at de tre modeller er næsten ens, hvorfor vi har valgt at gå videre med $q = -2$, idet dennes udgangsniveau ligger tættest på den observerede middelværdi af livvidde-ændringer i forhold til højde.



Figur 4.3: I figuren ses de tre modeller for det fraktionelle polynomium af første grad for $q = -2$, $q = -1$ og $q = -0,5$, for mænd.

Dermed fås for en mand, for den justerede model, følgende fraktionelle polynomier:

$$f_1(x_{\text{fed fisk}}, -2) = \beta_0 + \beta_1 x_{\text{fed fisk}}^{-2}$$

og

$$f_2(x_{\text{fed fisk}}, (0 \ 1)) = \beta_0 + \beta_1 \log(x_{\text{fed fisk}}) + x_{\text{fed fisk}}.$$

Vi har for alle fire modeller testet, hvorvidt FP2, FP1 eller en lineær model er at foretrække. Resultaterne af testene ses i Tabel 4.8. Vi har anvendt devians test og har valgt den metode der giver så kompleks en model som muligt indenfor denne ramme. I afsnit 10.2 uddybes proceduren for såvel den simpleste og den mest komplekse model. Det ses at for kvinderne er FP2 modellen signifikant bedre end FP1 modellen for både den justerede og ikke-justerede model. Derfor testes ikke yderligere og FP2 modellerne vælges. For mændene testes derimod både FP2 og FP1 modellen væk, men den lineære model beholdes.

Derfor modelleres mænds indtag af fed fisk som lineært, fremfor ved hjælp af et fraktionelt polynomium.

Model	Devians	d.f	q	Sammenligning	Dev. diff.	$P - vrdi$
Kvinder						
ikke-justeret						
FP2	445.297	4	3, 3	FP2 mod FP1	11	< 0,001
FP1	445.308	3	3	-	-	-
Justeret						
FP2	444.407	35	3, 3	FP2 mod FP1	6	< 0,05
FP1	444.413	34	3	-	-	-
Mænd						
ikke-justeret						
FP2	319.821,6	4	0, 0,5	FP2 mod FP1	2,6	0,1
FP1	319.824,2	3	2	-	-	-
Lineær	319.824,2	3	1	Lineær mod nul-model	319.824,2	< 0,001
Nul-model	0	0	-	-	-	-
Justeret						
FP2	319.062,8	32	0, 1	FP2 mod FP1	2,4	0,1
FP1	319.065,2	31	-1	FP1 mod lineær	0,6	0,5
Lineær	319.065,8	32	1	Lineær mod nul-model	329.065,8	< 0,001
Nul-model	0	0	-	-	-	-

Tabel 4.8: I tabellen ses resultaterne af devians testene for bestemmelse af grad af det fraktionelle polynomium.

Da vi for kvinder får, at det fraktionelle polynomium af anden grad, er den optimale grad for det fraktionelle polynomium, kan man overveje om det giver en endnu bedre model, at øge graden. Vi har valgt at stoppe ved anden grad, idet en højere grad end to sjældent er bedre [31] og fordi de metoder der på nuværende tidspunkt findes i STATA 11, til valg af den optimale potens og grad, ikke inkluderer brug af blandede modeller, hvorfor udvælgelsesprocessen er udført manuelt.

4.4 Valg af model

Vi har nu valgt antal knudepunkter for restringerede kubiske splines samt potens og grad af de fraktionelle polynomier. For at undersøge hvilke modeller der giver det bedste forhold mellem modellens evne til at forudsige fremtidige observationer og den kompleksitet, udregnes Akaike's informations kriterium for hver af modellerne. Resultaterne ses i Tabel 4.9. For yderligere information om Akaike's informations kriterium, se afsnit 8.1.1.

Model	d.f	LR	AIC
<u>Kvinder</u>			
Restringerede kubiske splines			
ikke-justeret ($k = 7$)	9	-445.307	-445.325
Justeret ($k = 7$)	40	-444.413	-444.493
Fraktionelle polynomier			
ikke-justeret ($p=(3 \ 3)$)	5	-423.575	-423.585
Justeret ($p=(3 \ 3)$)	36	-422.706	-422.778
Kategorisk			
ikke-justeret	7	-445.308	-445.322
Justeret	38	-444.418	-444.494
Lineær			
ikke-justeret	4	-445.304	-445.312
Justeret	35	-444.420	-444.490
<u>Mænd</u>			
Restringerede kubiske splines			
ikke-justeret ($k = 6$)	8	-319.819	-319.835
Justeret ($k = 6$)	36	-319.061	-319.133
Kategorisk			
ikke-justeret	7	-319.823	-319.837
Justeret	35	-319.065	-319.135
Lineær			
ikke-justeret	4	-319.824	-319.832
Justeret	32	-319.066	-319.130

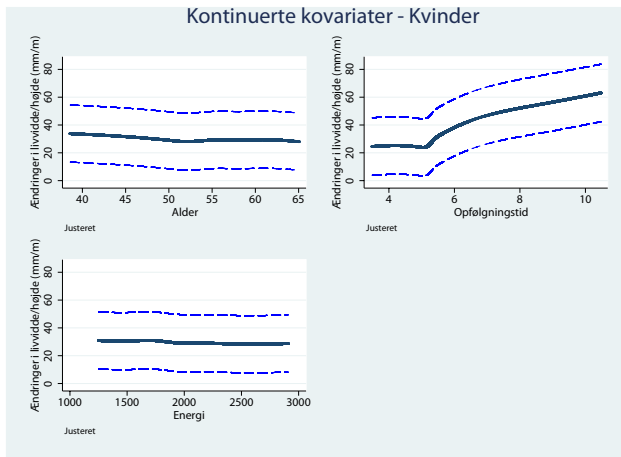
Tabel 4.9: I tabellen ses AIC for alle modellerne modelleret ved hjælp af blandede modeller. De værdier af AIC, der er markerede er de højeste for henholdsvis kvinder og mænd.

Af tabellen ses det, at den model der giver den højeste værdi af Akaike's informations kriterium for kvinder, er den justerede model, modelleret ved hjælp af det fraktionelle polynomium. Dermed har den endelige model for en kvinde k i land i og center j følgende forskrift:

$$\begin{aligned}
 f(x)_{ijk} = & \beta_0 + \beta_1 x_{\text{fed fisk}_k}^3 + \beta_2 \log(x_{\text{fed fisk}_k}) x_{\text{fed fisk}_k}^3 + \beta_3 \mathbb{I}_{\text{fed fisk}_{k_1}} \\
 & + \beta_4 \mathbb{I}_{\text{skole}_{k_2}} + \beta_5 \mathbb{I}_{\text{skole}_{k_3}} + \beta_6 \mathbb{I}_{\text{skole}_{k_4}} + \beta_7 \mathbb{I}_{\text{skole}_{k_5}} \\
 & + \beta_8 \mathbb{I}_{\text{ryge status}_{k_2}} + \beta_9 \mathbb{I}_{\text{ryge status}_{k_3}} + \beta_{10} \mathbb{I}_{\text{ryge status}_{k_4}} \\
 & + \beta_{11} \mathbb{I}_{\text{aktivitet}_{k_2}} + \beta_{12} \mathbb{I}_{\text{aktivitet}_{k_3}} + \beta_{13} \mathbb{I}_{\text{aktivitet}_{k_4}} \\
 & + \beta_{14} \mathbb{I}_{\text{menopausal}_{k_2}} + \beta_{15} \mathbb{I}_{\text{menopausal}_{k_3}} + \beta_{16} \mathbb{I}_{\text{menopausal}_{k_4}} \\
 & + \beta_{17} x_{\text{alder}_{k_1}} + \beta_{18} x_{\text{alder}_{k_2}} + \beta_{19} x_{\text{alder}_{k_3}} + \beta_{20} x_{\text{alder}_{k_4}} \\
 & + \beta_{21} x_{\text{alder}_{k_5}} + \beta_{22} x_{\text{alder}_{k_6}} + \beta_{23} x_{\text{opfølgningstid}_{k_1}} \\
 & + \beta_{24} x_{\text{opfølgningstid}_{k_2}} + \beta_{25} x_{\text{opfølgningstid}_{k_3}} + \beta_{26} x_{\text{opfølgningstid}_{k_4}} \\
 & + \beta_{27} x_{\text{opfølgningstid}_{k_5}} + \beta_{28} x_{\text{opfølgningstid}_{k_6}} + \beta_{29} x_{\text{energi}_{k_1}} \\
 & + \beta_{30} x_{\text{energi}_{k_2}} + \beta_{31} x_{\text{energi}_{k_3}} + \beta_{32} x_{\text{energi}_{k_4}} + \beta_{33} x_{\text{energi}_{k_5}} \\
 & + \beta_{34} x_{\text{energi}_{k_6}} + b_{\text{land}_i} + b_{\text{center}_j} + \varepsilon_{ijk},
 \end{aligned} \tag{4.1}$$

hvor β_0 er middelniveauet for lande, $b_{\text{land}_i} \sim \mathcal{N}(0, \sigma_{b_{\text{land}}}^2)$, $b_{\text{center}_j} \sim \mathcal{N}(0, \sigma_{b_{\text{center}}}^2)$ og $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

De ukendte parametre er blevet estimeret ved hjælp af restringeret maksimum likelihood estimation og ses i Tabel 4.10. Udfra koefficienterne kan vi, for de kategoriske variabler, se om de enkelte variabler har en indflydelse på ændringer i livvidde/højde. Som tidligere nævnt er første niveau for de kategoriske variabler reference niveau og de resterende niveauer kan ses i Tabel 3.2. Det ses at skolegang ikke har nogen signifikant effekt, mens der for rygning er en reducerende effekt ved at begynde at ryge og en øgende effekt at stoppe, ved fysisk aktivitet har det en reducerende effekt at være aktiv mens menopausalstatus ingen signifikant effekt har. Det ses ligeledes at størstedelen af variansen ikke forklares udfra land og center. For de kontinuerte variabler kan koefficienterne ikke tolkes separat, hvorfor vi har lavet en grafisk illustration af disse. Disse ses i Figur 4.4. I figuren ses det at alder har en let faldende tendens, dog ikke signifikant, opfølgningstid har et stort udsving ved over fem års opfølgning mens energiindtag ikke har nogen effekt.



Figur 4.4: I figuren ses de kontinuerte kovariater alder, opfølgningstid og energi modelleres som restringerede kubiske splines med syv knudepunkter, for kvinder.

4. OPSTILLING AF MODELLER

Variabel	Koefficient	Estimat	Konfidensinterval
Intercept	β_0	37,21	(10,84;63,59)
Fed fisk (FP)	β_1	4,88	(-1,31;11,07)
	β_2	22,20	(4,69;39,71)
Fed fisk (indikator)	β_3	0,30	(-1,85;2,44)
Skole (kategorisk)	β_4	-4,02	(-15,70;7,66)
	β_5	-6,62	(-18,32;5,08)
	β_6	-8,94	(-20,65;2,77)
	β_7	-10,50	(-22,22;1,23)
Rygning (kategorisk)	β_8	-0,95	(-2,08;0,18)
	β_9	-7,08	(-10,65;-3,51)
	β_{10}	14,78	(12,98;16,57)
Fysisk Aktivitet (kategorisk)	β_{11}	-0,51	(-1,86;0,84)
	β_{12}	-1,06	(-2,53;0,41)
	β_{13}	-2,88	(-4,37;-1,39)
Menopausalstatus (kategorisk)	β_{14}	-1,76	(-3,68;0,15)
	β_{14}	-0,70	(-2,51;1,11)
	β_{15}	-0,45	(-3,19;2,29)
Alder (RCS)	β_{16}	-0,24	(-0,47;0,00)
	β_{17}	-0,66	(-1,65;0,33)
	β_{18}	71,37	(3,51;139,23)
	β_{19}	-143,32	(-297,28;10,64)
	β_{20}	119,67	(-61,28;300,63)
	β_{21}	-72,00	(-210,08;66,07)
Opfølgningstid (RCS)	β_{22}	2,26	(-0,22;4,74)
	β_{23}	-26,00	(-59,25;7,25)
	β_{24}	11271,23	(6983,09;15559,36)
	β_{25}	-25000	(-36000;-15000)
	β_{26}	17715,60	(8459,90;26971,31)
	β_{27}	-3537,42	(-6475,07;-599,77)
Energi (RSC)	β_{28}	-0,00	(-0,01;0,00)
	β_{29}	0,06	(-0,06;0,18)
	β_{30}	-0,50	(-1,23;0,24)
	β_{31}	1,12	(-0,42;2,66)
	β_{32}	-1,09	(-2,74;0,56)
	β_{33}	0,52	(-0,44;1,48)

Tabel 4.10: I tabellen ses regressionskoefficienterne fundet ved restringeret maksimum likelihood estimation, for modellen for kvinder. Varianskomponenterne $b_{\text{land}} \sim \mathcal{N}(0,202)$, $b_{\text{center}} \sim \mathcal{N}(0,123)$ og $\varepsilon_{ijk} \sim \mathcal{N}(0,2.067)$.

For mændene er den model med den højeste værdi, af akaike's informations kriterium, den justerede lineære model. Dermed har den endelige model for en mand m i land i og center j forskriften:

$$\begin{aligned}
f(x)_{ijm} = & \beta_0 + \beta_1 x_{\text{fed fisk}_m} + \beta_2 \mathbb{I}_{\text{fed fisk}_{m_1}} \\
& + \beta_3 \mathbb{I}_{\text{skole}_{m_2}} + \beta_4 \mathbb{I}_{\text{skole}_{m_3}} + \beta_5 \mathbb{I}_{\text{skole}_{m_4}} + \beta_6 \mathbb{I}_{\text{skole}_{m_5}} \\
& + \beta_7 \mathbb{I}_{\text{ryge status}_{m_2}} + \beta_8 \mathbb{I}_{\text{ryge status}_{m_3}} + \beta_9 \mathbb{I}_{\text{ryge status}_{m_4}} \\
& + \beta_{10} \mathbb{I}_{\text{aktivitet}_{m_2}} + \beta_{11} \mathbb{I}_{\text{aktivitet}_{m_3}} + \beta_{12} \mathbb{I}_{\text{aktivitet}_{m_4}} \\
& + \beta_{13} x_{\text{alder}_{m_1}} + \beta_{14} x_{\text{alder}_{m_2}} + \beta_{15} x_{\text{alder}_{m_3}} + \beta_{16} x_{\text{alder}_{m_4}} \\
& + \beta_{17} x_{\text{alder}_{m_5}} + \beta_{18} x_{\text{alder}_{m_6}} + \beta_{18} x_{\text{opfølgningstid}_{m_1}} \\
& + \beta_{20} x_{\text{opfølgningstid}_{m_2}} + \beta_{21} x_{\text{opfølgningstid}_{m_3}} + \beta_{22} x_{\text{opfølgningstid}_{m_4}} \\
& + \beta_{23} x_{\text{opfølgningstid}_{m_5}} + \beta_{24} x_{\text{opfølgningstid}_{m_6}} + \beta_{25} x_{\text{energi}_{m_1}} \\
& + \beta_{26} x_{\text{energi}_{m_2}} + \beta_{27} x_{\text{energi}_{m_3}} + \beta_{28} x_{\text{energi}_{m_4}} \\
& + \beta_{29} x_{\text{energi}_{m_5}} + \beta_{30} x_{\text{energi}_{m_6}} + b_{\text{land}_i} + b_{\text{center}_j} + \varepsilon_{ijm},
\end{aligned} \tag{4.2}$$

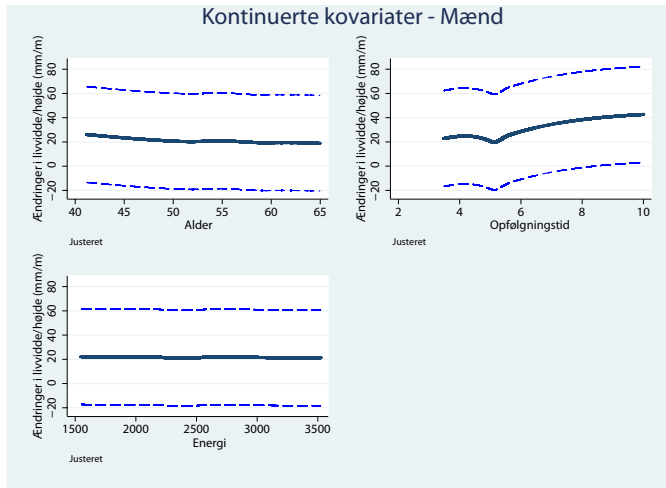
hvor β_o er middelniveauet landene, $b_{\text{land}_i} \sim \mathcal{N}(0, \sigma_{b_{\text{land}}}^2)$, $b_{\text{center}_j} \sim \mathcal{N}(0, \sigma_{b_{\text{center}}}^2)$ og $\varepsilon_{ijm} \sim \mathcal{N}(0, \sigma^2)$.

De ukendte parametre er blevet estimeret ved hjælp af restringeret maksimum likelihood estimation og ses i Tabel 4.11. Det ses at skolegang ikke har nogen signifikant effekt, mens der for rygning er en reducerende effekt ved at begynde at ryge og en øgende effekt at stoppe mens fysisk aktivitet ingen signifikant effekt har. For de kontinuerte variabler kan koefficienterne ikke tolkes separat, hvorfor vi har lavet en grafisk illustration af disse. Det ses ligeledes at størstedelen af variansen ikke forklares ud fra land og center. Disse ses i Figur 4.5. I figuren ses det at alder har en let faldende tendens, dog ikke signifikant, opfølgningstid har et stort udsving ved over fem års opfølgning mens energiindtag ikke har nogen effekt.

4. OPSTILLING AF MODELLER

Variabel	Koefficient	Estimat	Konfidensinterval
Intercept	β_0	23,78	(-18,21;65,78)
Fed fisk (lineær)	β_1	-0,19	(-0,47;0,08)
Fed fisk (indikator)	β_2	-0,94	(-33,72;43,89)
Skole (kategorisk)	β_3	5,08	(-15,70;7,66)
	β_4	6,51	(-32,30;45,32)
	β_5	7,32	(-31,49;46,13)
	β_6	6,81	(-31,99;45,62)
	β_7	0,33	(-0,61;1,27)
Rygning (kategorisk)	β_8	-8,29	(-10,82;-5,75)
	β_9	12,85	(11,39;14,30)
	β_{10}	0,58	(-0,68;1,84)
Fysisk Aktivitet (kategorisk)	β_{11}	1,18	(-0,13;2,48)
	β_{12}	1,09	(-0,19;2,37)
	β_{13}	-0,84	(-1,06;-0,63)
Alder (RCS)	β_{14}	0,83	(-0,20;1,87)
	β_{15}	34,81	(-24,17;93,79)
	β_{16}	-99,05	(-228,12;30,03)
	β_{17}	122,41	(-16,25;261,07)
	β_{18}	-87,54	(-190,44;15,36)
Opfølgningstid (RCS)	β_{19}	6,89	(4,69;9,08)
	β_{20}	-76,43	(-105,85;-47,01)
	β_{21}	9214,96	(3101,19;15328,73)
	β_{22}	-16000	(-38000;4986,27)
	β_{23}	3762,53	(19000;26398,32)
	β_{24}	4388,92	(-4563,49;13341,33)
Energi (RSC)	β_{25}	-0,00	(-0,01;0,00)
	β_{26}	0,02	(-0,05;0,10)
	β_{27}	-0,22	(-0,74;0,29)
	β_{28}	0,64	(-0,45;1,73)
	β_{29}	-0,85	(-2,02;0,32)
	β_{30}	0,52	(-0,16;1,24)

Tabel 4.11: I tabellen ses regressionskoefficienterne fundet ved restringeret maksimum likelihood estimation, for modellen for kvinder. Varianskomponenterne $b_{\text{land}_i} \sim \mathcal{N}(0,3)$, $b_{\text{center}_j} \sim \mathcal{N}(0,82)$ og $\varepsilon_{ijm} \sim \mathcal{N}(0,1.172)$.



Figur 4.5: I figuren ses de kontinuerte kovariater alder, opfølgningstid og energi modelleres som restringerede kubiske splines med syv knudepunkter, for mænd.

Antallet af parametre brugt i modellerne er henholdsvis 36 for kvinder og 32 for mænd, hvilket er langt færre end troværdighedsmålet tillader. Vi benytter derfor resultaterne af disse modeller til, at forklare sammenhængen mellem indtag af fed fisk og ændringer i livvidde i forhold til højde. Vi betragter ligeledes resultaterne fra nogle af de andre modeller for, at se hvilken betydning det har for resultaterne, hvis vi havde valgt en anden model.

RESULTATER

Hovedresultaterne for analyserne findes udfra de optimale modeller, valgt i kapitel 4. Udover hovedresultaterne, sammenlignes resultaterne fra udvalgte modeller for, at vise hvilken betydning det har at vælge forskellige modeller. Alle illustrationer, af fraktionelle polynomier og restringerede kubiske splines, i dette afsnit, er vist mellem 5 procent og 95 procent percentilerne af fed fisk.

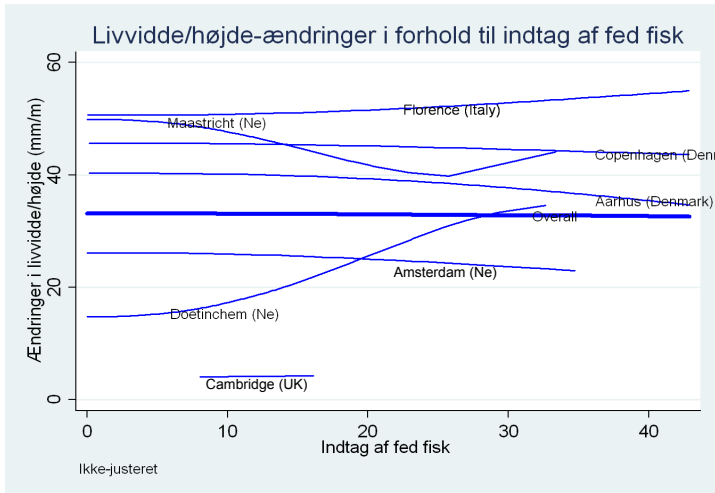
5.1 Resultater - Kvinder

Den optimale model for kvinder, var den model hvor eksponeringen var modelleret som et fraktionelt polynomium af anden grad med $\mathbf{q} = (3\ 3)$, justeret for de potentielle confoundere: socialstatus, rygeændringer, fysisk aktivitet alder, opfølgningsperiode, energi og menopausalstatus. Til at vurdere hvorvidt det er fornuftigt, at anvende den blandede model til, at modellere sammenhængen mellem indtag af fed fisk og ændringer i livvidde/højde ønskes, at lave en multivariat metaanalyse med fast effekt for den tilsvarende model, jævnfør kapitel 9.

På grund af perfekt prædiktion, jævnfør afsnit 9.5, er det ikke muligt, at lave metaanalysen for de samme justeringer, som der anvendes ved den blandede model. Programmet *mvmeta_make*, skrevet af Ian White, forsøger at omgås perfekt prædiktion ved at ændre to værdier for hver kovariat, til henholdsvis gennemsnittet af kovariaten i det specifikke studie plus og minus standardafvigelsen. Denne form for regulering medfører at perfekt prædiktion kan omgås i de kategoriske variabler. Vi justere for de kontinuerte forklarende variabler, ved hjælp af restringerede kubiske splines, som også opdeler data i grupper. I STATA 11 generes $k - 1$ nye variabler som anvendes til at justere for den givne variabel. Hvis der opstår perfekt prædiktion for en af disse nye variabler, i et eller flere af studierne, kan Ian Whites regulerings ikke anvendes, idet spredningen giver nul og metoden ændre derved ikke på variablen.

For at kunne lave en justeret multivariat metaanalyse helt uden perfekt prædiktion, skal indikatorvariablen for fed fisk og folk der ikke har fuldført folkeskolen, fjernes fra analyserne samtidig med, at de kontinuerte variabler skal justeres for som kategoriske eksempelvis inddelt i kvintiler. Resultaterne fra de to metoder kan derfor ikke umiddelbart sammenlignes, idet den blandede model er justeret på én måde, mens metaanalysen er justeret på en anden måde. Derfor har vi valgt kun at anvende de ikke-justerede metaanalyser. For de ikke-justerede metaanalyser opstår stadig perfekt prædiktion for indikatorvariablen for fed fisk, idet nogle af centrene ikke har nul-indtagere med. Da der her er tale om en kategorisk variabel kan perfekt prædiktion her omgås ved hjælp af Ian Whites regulering. Dermed er der justeret ens i den ikke-justerede blandede model og den ikke-justerede metaanalyse, hvorfor vi kan sammenligne resultaterne af de to modeller. Da vi anvender metaanalysen til, at vurdere hvorvidt sammenhængen mellem eksponeringen og responsvariablen er ensartet, på tværs af studierne, og ikke til det endelige resultat, mener vi godt, at vi kan anvende den ikke-justerede multivariate metaanalyse til, at vurdere om studierne er homogene, samt anvende denne viden for den justerede blandede model.

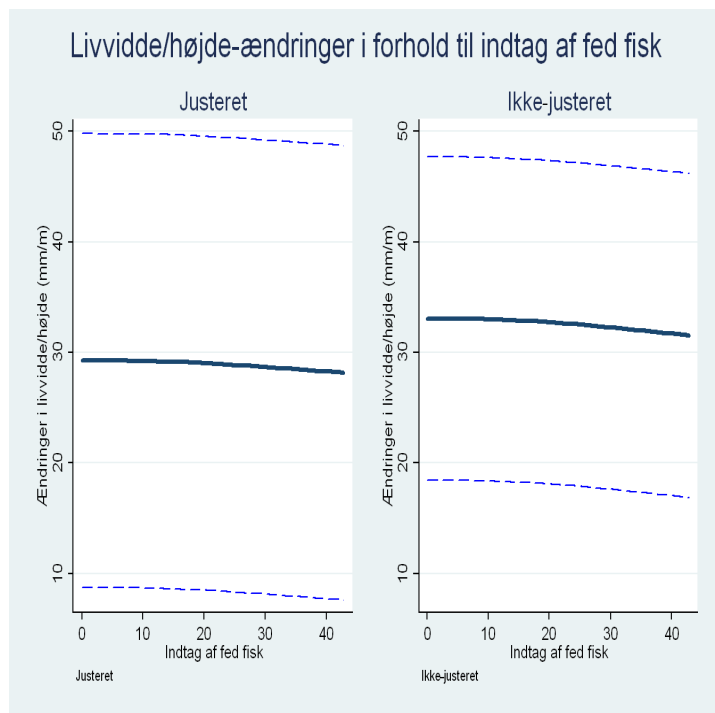
Resultatet af metaanalysen ses i Figur 5.1. Det ses af den ikke-justerede metaanalyse, at studierne København, Århus, Maastricht, Cambridge og Amsterdam har en let faldende tendens, mens studierne Firenze og Doetinchem viser en tendens til, at indtag af fed fisk har en øgende effekt på ændringen af livvidde i forhold til højde. De to hollandske studier Maastricht og Doetinchem har den tydeligste ikke-lineære sammenhæng mellem indtag af fed fisk og ændringer i livvidde/højde, de viser dog modsatrettede tendenser. Vi vurdere udfra sammenligningen af studierne i kapitel 3 og den multivariate metaanalyse for den ikke-justerede model, at studierne er ensartede nok til, at analysen foretaget med en blandet model giver mening.



Figur 5.1: I figuren ses resultatet den ikke-justerede multivariate metaanalyse for sammenhængen mellem livvidde ændringer /højde og indtag af fed fisk, for kvinder.

Resultatet for middelniveauet af den blandede model, for den ikke-justerede og justerede model, ses i Figur 5.2. Det ses, at forskellen mellem den justerede og den ikke-justerede model er, at udgangsniveauet for ændringer i livvidde i forhold til højde rykkes fra lidt over 30 i den ikke-justerede model, til lidt under 30 i den justerede model. Konfidensintervallet for den justerede model bliver ligeledes lidt bredere. De centerspecifikke resultater ses i appendix B. De centerspecifikke resultater er ikke-justerede, idet de kendte værktøjer til at illustrere justerede modeller i STATA 11, ikke virker sammen med pakkerne for blandede modeller, hvor de centerspecifikke niveauer prædikteres.

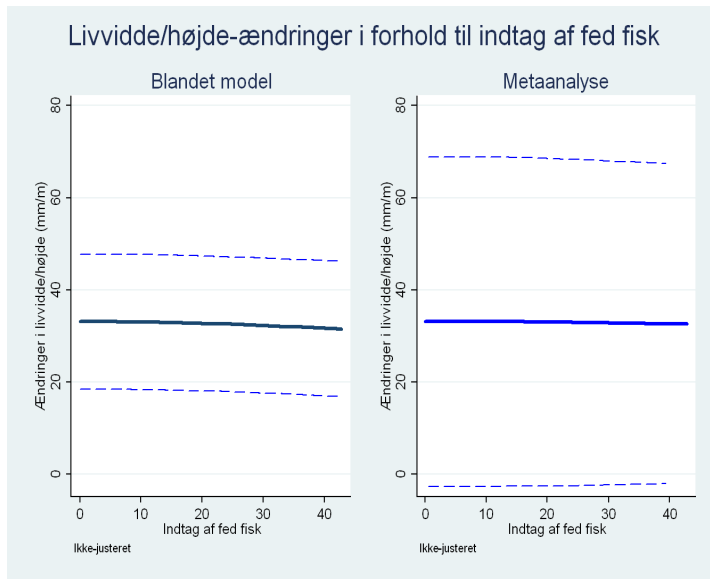
Vores hovedresultat for kvinderne er den justerede blandede model i Figur 5.2. Resultatet af denne analyse viser en let faldende tendens i forøgelsen af livvidde i forhold til højde. Konfidensbåndet er så bredt, at den mindste værdi af det øvre konfidensbånd er større end den største værdi af regressionslinien og den største værdi af det nedre konfidensbånd er mindre end den mindste værdi af regressionslinien, hvilket betyder at den reducerende effekt ikke er signifikant.



Figur 5.2: I figuren ses resultatet af den blandede model for sammenhængen mellem livvidde ændringer /højde og indtag af fed fisk modelleret med et fraktionelt polynomium med $\mathbf{q} = (q_1 \ q_2) = (3 \ 3)$ henholdsvis justeret og ikke-justeret, for kvinder.

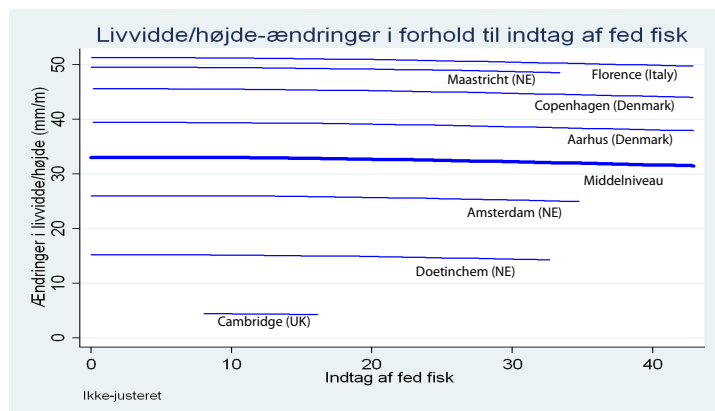
Som tidligere nævnt, kan resultaterne af metaanalysen ikke sammenlignes med resultatet af den blandede model, idet der ikke er justeret for potentielle confoundere i metaanalysen. Derfor sammenlignes med den ikke-justerede blandede model. Det vægtede gennemsnit med tilhørende konfidensbånd fra den multivariate metaanalyse samt resultatet for middelniveauet med tilhørende konfidensbånd for den blandede model, ses i Figur 5.3. Det ses at estimatet for kurven er stort set den samme, men konfidensintervallet for den multivariate metaanalyse er en del bredere end for den multivariate metaanalyse. Det brede konfidensbånd for metaanalysen kan skyldes, at de enkelte centre har meget forskellige udgangsniveauer. I Figur 5.1 så vi at

centrenes udgangsniveau lå fra cirka 5 (mm/m) - 50 (mm/m). For at se hvor meget centrene fordeler sig ved brug af den blandede model har vi indtegnet de, ikke-justerede, centerspecifikke grafer sammen med grafen for middelniveauet, hvilket ses i Figur 5.4. Her ses det ligeledes at centrene spreder sig fra omkring 5 (mm/m) til over 50 (mm/m).



Figur 5.3: I figuren ses resultatet af henholdsvis den blandede model og den multivariate metaanalyse for sammenhængen mellem livvidde ændringer delt med højde og indtag af fed fisk modelleret med et fraktionelt polynomium med $\mathbf{q} = (q_1 \ q_2) = (3 \ 3)$, for kvinder.

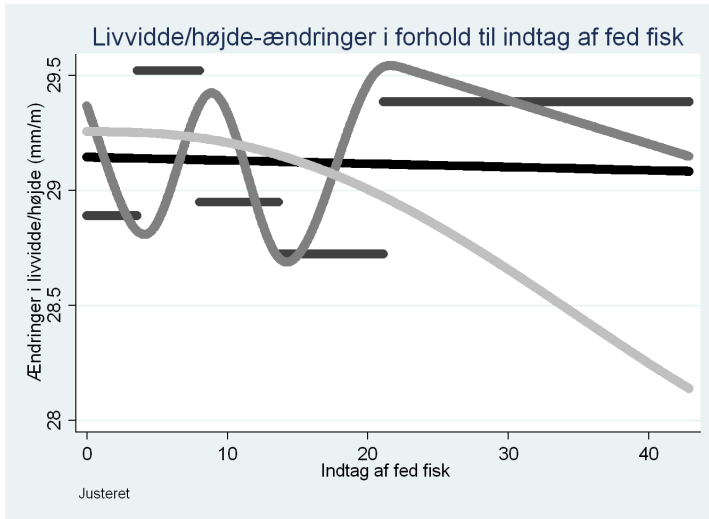
På grund af det meget brede konfidensbånd, kan formen af det fraktionelle polynomium ikke vurderes. I Figur 5.5 har vi derfor indtegnet de justerede blandede modeller for middelniveauet for center, hvor eksponeringsvariablen, fed fisk, er modelleret som henholdsvis lineære, kategoriske, restringeret kubisk spline og fraktionelt polynomium. Det ses, at i modellen hvor en restringeret kubisk spline er brugt, fås en tre aftagende perioder og tre voksende perioder. Mellem nul og fem gram reduceres ændringen i livvidden, mellem fem og ti stiger ændringen igen hvilket fortætter tre svingninger. Den lineære model og den hvor



Figur 5.4: I figuren ses de centerspecifikke grafer for den ikke-justerede blandede model, for kvinder.

fraktionelle polynomium er brugt har en konstant faldene effekt. Den faldende effekt er størst ved det fraktionelle polynomium. Kvintilerne underbygger formen af den restringerede kubiske spline og det er derfor vigtigt, at vurdere om man vil anvende den mest datanære model, eller en model der fanger tendensen men ikke ligger helt tæt op ad data. Vi har valgt den model, som gav det bedste forhold mellem modellens evne til at forudsige fremtidige observationer og modellens kompleksitet. Vi er af den generelle mening, at fraktionelle polynomier er at foretrække til modellering af eksponeringsvariabler, hvor man har en formodning om, at sammenhængen ikke har store udsving, mens restringerede kubiske splines er velegnet til at fjerne confounding effekt og modellering af eksponeringsvariabler, hvor man ingen forhånds antagelser har.

Alle modellerne som afbildedes i Figur 5.5 falder indenfor konfidensbåndet for modellen for det fraktionelle polynomium, som i grove træk strækker sig fra 10 (mm/m) - 50 (mm/m). Effekten af indtag af fed fisk ville derfor ikke være signifikant ligegyldig hvilken model vi have valgt. Havde effekten vi har fundet været signifikant, burde vi igen overveje om resultatet havde en klinisk relevans. I modellen, hvor indtag af fed fisk er modelleret som et fraktionelt polynomium, er forskellen fra et minimalt indtag af fed fisk, til et indtag på over



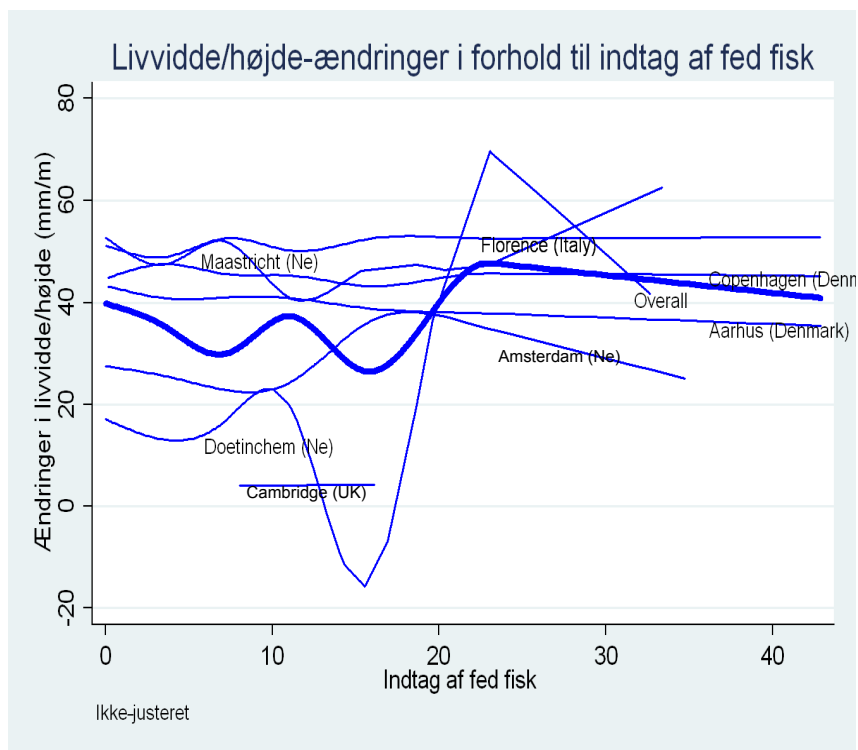
Figur 5.5: I figuren ses den model hvor sammenhængen mellem indtag af fed fisk responsvariablen, ændringer i livvidde/højde, er modelleret som lineære, kategoriske, med restringerede kubiske splines og med et fraktionelt polynomium.

40 gram per dag, cirka 1,2 (mm/m). Dette vil for en kvinde med højden 1,67 meter betyde en reducere af livvidde ændringen på 2 millimeter, i løbet af opfølgingsperioden.

Vi har valgt, at anvende modellen, hvor det fraktionelle polynomium er brugt til, at modellere eksponeringsvariablen. Havde vi fået at modellen, hvor restringerede kubiske splines var brugt, gav den bedste model, skulle vi ligeledes have foretaget en multivariat metaanalyse for denne tilsvarende model, for at kunne vurdere om centrene var ensartede.

Restringerede kubisk splines er mere frie end fraktionelle polynomier og vi forventede derfor ikke a priori at en multivariat metaanalyse for restringerede kubiske splines vil give fortløkkelige resultater. Vi har derfor lavet en multivariat metaanalyse for den ikke-justerede model, hvor syv knudepunkter er anvendt, for at undersøge om vores forventning var begrundet. Analysen ses i Figur 5.6. Det ses at studierne Maastricht, Amsterdam og Doetinchem har tendens

til, at lave store udsving, mens de resterende studier har mindre eller slet ingen udsving. Flere af studierne viser samme tendens til først at aftage, efterfulgt af to bakkedale. Det er sværere at vurdere, hvorvidt studierne kan betragtes som et samlet studie. Ved kun at betragte plottet, ville vi vurdere at København, Århus og Florence opfører sig rimelig ens i forhold til sammenhængen mellem indtag af fed fisk og ændringer i livvidde/højde. Maastricht, Doetinchem og Amsterdam har næsten samme form endog lidt forskudt, mens Cambridge falder helt udenfor. At Cambridge falder udenfor, kan skyldes, at udfaldsrummet for indtag af fed fisk for Cambridge er mindre end for de andre studier.



Figur 5.6: I figuren ses den multivariate metaanalyse for sammenhængen mellem livvidde ændringer /højde og indtag af fed fisk modelleret med en restriungeret kubisk spline med $k = 7$ for kvinder.

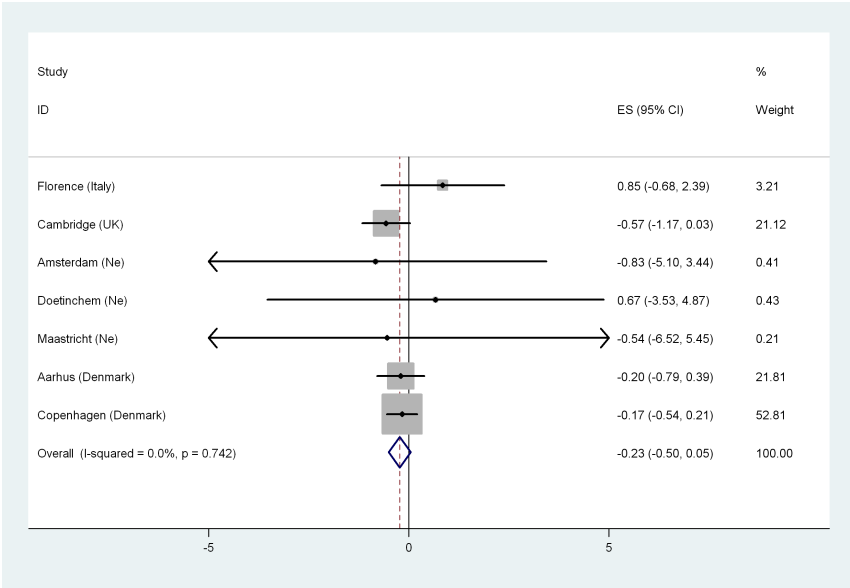
Vores konklusion er, at det er sværere at vurdere om tendensen for de enkelte studier er ens for modellen, hvor eksponeringen er modelleret som en restringeret kubisk spline, fremfor som et fraktionelt polynomium. I netop dette tilfælde hvor kun syv studier er betragtet, mener vi at kunne se en overordnet fælles tendens, men mener ligeledes, at det vil blive meget besværligt at vurdere, hvis antallet af studier øges.

5.2 Resultater - Mænd

Den optimale model for mænd var den lineære model justeret for de potentielle confoundere: socialstatus, rygeændringer, fysisk aktivitet alder, opfølgningsperiode og energi. Til at vurdere hvorvidt det er fornuftigt at anvende den blandede model til, at modellere sammenhængen imellem indtage af fed fisk og ændringer i livvidde/højde, testes for heterogenitet og der udføres en metaanalyse for den tilsvarende model, jævnfør kapitel 9. Hvis testet for heterogenitet er signifikant laves en metaanalyse med tilfældig effekt ellers laves metaanalysen med fast effekt.

Testet for heterogenitet gav en p-værdi på $p = 0,907$ hvilket betyder at nulhypotesen, at effekten af indtag af fed fisk i centrene er ens, ikke kan afvises. Metaanalysen laves derfor med fast effekt. Resultaterne angives som 10 gram forøget indtag af fed fisk, idet effekten af et grams forøgelse først ses på anden decimal. Resultatet af metaanalysen ses i Figur 5.7. Det ses at alle centrene ligger tæt omkring linjen gennem nul. Nogle ligger lidt over mens andre ligger lidt under. Fælles for alle studierne er dog, at konfidensintervallet går ind over nul. Vi vurderer ud fra udvælgelseskriterierne i kapitel 3, heterogenitets testet og metaanalysen, at de syv centre Firenze, Cambridge, Amsterdam, Doetinchem, Maastricht, Århus og København, kan opfattes som et studie, når sammenhængen mellem indtag af fed fisk ændringer i livvidde/højde for mænd, betragtes som lineær.

Resultaterne for den lineære blandede model, ses for den ikke-justerede og justerede model i Tabel 5.1. Estimerne er fundet ved hjælp af restringeret maksimum likelihood estimation. Konklusionen af



Figur 5.7: I figuren ses den justerede metaanalysen for sammenhængen mellem forøgelse af indtaget af fed fisk med 10 gram og ændringer i livvidde/højde (mm/m) for mænd.

analysen er, at der ses en ikke signifikant tendens til at mænds forhold mellem livvidde ændringer og højde reduceres, ved øget indtag af fed fisk. Effekten er ikke signifikant og hvis den havde været det, kunne det diskuteres, hvorvidt effekten var klinisk relevant. Reduceringen af livvidde/højde er 0,19 (*mm/m*), i løbet af ens opfølgningsperiode, ved en forøgelse af det daglige indtag af fed fisk med 10 gram per dag. For en mand med højden 1,8 meter betyder det en livvidde reducereing på 0,34 millimeter.

Model	Estimat	Konfidensinterval
ikke-justeret	-0,21	(-0,48; 0,06)
Justeret	-0,19	(-0,47; 0,08)

Tabel 5.1: I tabellen ses resultaterne af henholdsvis den ikke-justerede og justerede lineære model for mænd

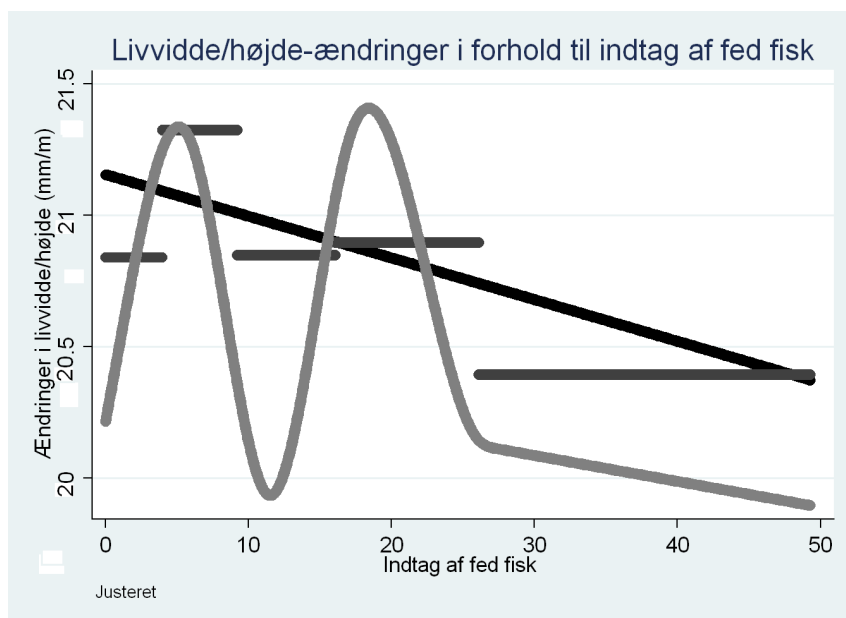
Resultaterne for den ikke-justerede model er medtaget for at se, hvilken effekt det har på parameter estimatet for def fisk og det tilhørende konfidensinterval, at justere for de variabler, vi har vurderet er forklarende variabler. Det ses, at resultaterne er rykket en lille smule tættere på nul, hvilket viser, at de variable vi har valgt at justere for forklarer noget af ændringen af livvidde i forhold til højde. Konfidensintervallet bliver en lille smule bredere, men rykkes tilsvarende en smule mod højre. I kapitel 4 Tabel 4.11 og Figur 4.5 betragtede vi, hvilken effekt, de enkelte forklarende variabler, har på livvidde-ændringer i forhold til højde.

Alle centrene har den samme hælningskoefficient, men har forskellige udgangsniveauer. De centerspecifikke modeller kan ses i appendix C.

Sammenlignes med resultatet af metaanalysen i Figur 5.7, hvor vi fik at hælningskoefficienten for 10 grams forøget indtag af fisk var -0,23 (-0,50;0,05), ses det, at resultatet af metaanalysen og resultatet af den blandede model giver samme konklusion, men ikke præcis samme estimater.

I kapitel 4 udvalgte vi den lineære model, som den model der gav det bedste forhold mellem modellens evne til at forudsige fremtidige observationer og kompleksitet. For at se hvilken virkning det ville have haft at modellere problemstillingen med andre modeller, har vi estimeret effekten for modellerne hvor eksponeringen er medtaget som kategorisk inddelt i kvintiler og som restringerede kubiske splines, med seks knudepunkter, og indtegnet dem i Figur 5.8. Det ses at alle tre modeller har en nedadgående tendens således, at den forøgelse der er i livvidden i forhold til højden, reduceres med indtaget af fed fisk. Den lineære og kategoriske model følger hinanden rimeligt godt, men en del af informationen, vedrørende sammenhængen, går tabt i springene mellem kategorier. Havde sammenhængen været ikke-lineær, ville den kategoriske model have givet mere information end den lineære. Modellen modelleret ved hjælp af en restringeret kubisk spline har to store udsving, hvilket fortolkningsmæssigt ville betyde at man ved indtag af lidt fisk forøgede ændringerne i livvidden, i en negativ retning, for derefter igen at reducere ændringen indtil

cirka 10 gram fed fisk og tilsvarende for den anden pukkel. Dette viser hvor vigtigt valget af model er. I dette tilfælde giver den lineære model den bedste forhold mellem modellen evne til at forudsige fremtidige observationer i forhold til kompleksitet, men havde sammenhængen i virkeligheden været ikke-lineær ville meget information være gået til spilde ved, at tvinge en lineær model ned over data.



Figur 5.8: I figuren ses den model hvor sammenhængen mellem indtag af fed fisk responsvariablen, ændringer i livvidde/højde, er modelleret som lineære, kategoriske og med restringerede kubiske splines for mænd.

For at se, hvilken effekt, ekstreme værdier af indtag af fed fisk har, på ændringer i livvidden i forhold til højden har vi ligeledes lavet sensitivitetsanalyser, hvor vi fjernet alle deltagere med indtag over 100 gram fed fisk om dagen. Analyserne blev derefter lavet på samme vis, som for det fulde datasæt. Vi har valgt ikke at tage resultaterne fra sensitivitetsanalyserne med i rapporten, idet de viste det samme som for det fulde datasæt.

KONKLUSION

Formålet med dette speciale er, at undersøge hypotesen vedrørende ændringer i livvidde/højde i forhold til indtag af fed fisk, sammenholde metoderne blandede modeller og metaanalyse ved, at belyse teoretiske såvel som tekniske fordele og ulemper samt sammenligne metoderne restringerede kubiske splines med fraktionelle polynomier, praktisk såvel som teoretisk. Sammenligningen af de teoretiske metoder er blev foretaget i lyset af problemstillingen vedrørende ændringer i livvidde/højde i forhold til indtag af fed fisk. Derudover ville vi udvikle programmer til STATA 11, som muliggjorde grafiske præsentation af multivariat metaanalyse for restringerede kubiske splines og fraktionelle polynomier.

Ved at se på den observerede sammenhæng mellem livvidde-ændringer og højde i forhold til indtag af fed fisk i Figur 3.3 side 24 så vi, at der umiddelbart så ud til, at være en stationær ændring i livvidde/højde omkring et middelniveau, for hvert center, men at det ikke ændrede sig i forhold til indtag af fed fisk. Efter at have modelleret sammenhængen med de modeller vi, udfra vores kriterier fandt bedst, kom vi frem til samme konklusion.

Vi stratificerede analyserne på køn, idet vi forventede at se forskellige tendenser for kvinder og mænd. For de to køn blev der valgt forskellige modelleringsformer for indtag af fed fisk, idet de opstillede udvælgelseskriterier gav en ikke-lineær model for kvinder og en lineær model for mænd. På trods af den forventede forskel i retning af effekten, for indtag af fed fisk på ændringer i livvidden/højde, gav de to modeller samme konklusion, at der ingen signifikant effekt var på livvidde-ændringer/højde af at indtage fed fisk. Ved at se på varianskomponenterne for modellerne, i Tabel 4.10 for kvinder og Tabel 4.11 for mænd, så vi ligeledes at den største variation i livvidde-ændringer/højde ikke skyldtes center eller lande effekt, men indgik i residualvariationen.

Selvom vi ikke har vist en positiv effekt af indtag af fed fisk, har vi fået en større forståelse af årsagsmekanismer for fedme. Vi ved nu at indtag af fed fisk ikke har nogen signifikant effekt på ændringer i abdominal fedme, men vi har set en meget svag tendens til en positiv effekt, det vil sige en negativ ændring af livvidde/højde. Det kan være at effekten bliver tydeligere, hvis et bedre mål for indtag af fed fisk anvendes. Det nuværende mål for indtag af fed fisk er blevet målt ud fra et kostfrekvensspørgeskema som spørger til deltagerens kostindtag et år tilbage i tiden, hvilket er et meget usikkert mål. Et mere præcist mål kunne være mængden af ω -3 fedtsyrer i kroppen.

Derudover fik vi mulighed for, at udforske en række problemstillinger vedrørende valg af studiepopulation og opstilling af model. Disse problemstillinger kan opstå når teoretiske metoder anvendes i praksis, hvorfor vi ikke kunne have fået indsigt i disse uden, at have betragtet hypotesen vedrørende sammenhængen mellem indtag af fed fisk og ændringer i livvidde i forhold til højde.

Udover at betragte hovedhypotesen angående indtag af fed fisk og ændringer i livvidde i forhold til højde, var vores formål ligeledes at sammenligne de to metoder multivariat metaanalyse og blandede modeller. Vi så for den fortrukne model for mænd, at univariat metaanalyse gav samme konklusion som den blandede model. For kvinderne blev det nødvendig at anvende multivariat metaanalyse, idet eksponeringen blev modelleres som et fraktionelt polynomium. Ved anvendelse af multivariat metaanalyse, stødte vi ind i problemer med perfekt prædiktion, som følger af at studiernes udfaldsrum for de enkelte justeringsvariabler var forskellige. Ved at fjerne de variabler som resulterede i perfekt prædiktion, kunne vi lave en analyse hvor der var justeret for nogle af variablerne, men besluttede at holde os til den ikke-justerede model, idet vi derved kunne sammenligne resultaterne med resultaterne fra den ikke-justerede blandede model. I tilfælde af at man vil justere for potentielle confoundere, som ikke har samme udfaldsrum i alle de implicerede studier, bør man derfor være opmærksom på, at det regneteknisk kan give problemer, at vælge multivariat metaanalyse fremfor blandede modeller.

De blandede modeller gav ligeledes regnetekniske problemer, da de centerspecifikke resultater skulle plottes. Da vores hovedvægt ligger på det samlede estimat, som det var muligt at illustrere grafisk, fandt vi ikke denne forhindring lige så graverende som ikke at kunne justere som ønsket i de multivariate metaanalyser. Selvom det ikke er hovedresultatet, vil det være en fordel ligeledes, at kunne se de centerspecifikke resultater for justerede modeller.

Vores erfaring har vist os, at pakkerne til multivariat metaanalyse og grafisk illustration af blandede modeller i STATA 11, ikke er så udviklede, som man kunne ønske sig. Udover sammenligningen af anvendelsen af de to metoder foretages, i kapitel 7 og 9 en gennemgang af de to metoder og en mere teoretisk sammenligning fortages.

For at det er fornuftigt at anvende en blandet model, skal der være homogenitet mellem studierne. Til at vurdere dette har vi betragtet studiepopulationen i de enkelte studier, testet for homogenitet i det univariate tilfælde og betragtet de tilsvarende metaanalyser i det univariate såvel som multivariate tilfælde. Vi mener denne sammenligning er meget vigtig og mener ikke de forskellige delelementer, til vurderingen af homogenitet, kan stå alene.

Vi ønskede ligeledes at sammenligne metoderne restringerede kubiske splines og fraktionelle polynomier. De to metoder er begge metoder til ikke-lineær modellering, men modellerer ikke-lineariteten forskelligt. For resultaterne for indtag af fed fisk i forhold til ændringer i livvidde/højde, så vi at de to metoder afspejler samme tendens, men at forløbet af kurverne er meget forskelligt. Restringerede kubiske splines giver et godt billede af, hvordan sammenhængen indenfor et bestemt område i et bestemt datasæt er, hvor de fraktionelle polynomier stadig afhænger af hvilket datasæt modellen fittes for, men er knap så datanære. Ved multivariat metaanalyse var det nemmere, at vurdere om sammenhængen mellem eksponeringsvariablen og responsvariablen var ensartet for fraktionelle polynomier, idet disse ikke har så store udsving som restringerede kubiske splines. Ved brug af fraktionelle polynomier i forbindelse med blandede modeller, var vi nød til manuelt, at vælge den optimale potens og grad, hvilket

er en ulempe ved de nuværende pakker til fraktionelle polynomier i STATA 11. I tilfælde af, at problemstillingen ikke er lineær, er begge metoder at foretrække i forhold til en lineær eller kategorisk model. Diskussionen af om fraktionelle polynomier eller restringerede kubiske splines er at foretrække, fortsættes i kapitel 10, efter de to metoder er blevet gennemgået.

Alt efter hvordan eksponeringsvariablen blev modelleret fik vi forskellige fortolkninger af sammenhængen mellem indtag af fed fisk og livvidde-ændringer/højde. I vores tilfælde viste alle modellerne en let nedadgående tendens som ikke var signifikant, hvorfor valget af modelleringsform ikke ville have haft stor betydning for vores resultater. I tilfælde af at der havde været en effekt af at indtage fed fisk viser Figur 5.5 og 5.8 at fortolkningen over mængden af fisk er meget forskellig for modelleringsformerne lineær, kategorisk, restringerede kubiske splines og fraktionelle polynomier.

Da der ikke tidligere har fandtes programmer til grafisk at illustrere fraktionelle polynomier og restringerede kubiske splines efter multivariat metaanalyse, var et af delmålene for dette speciale, at programmere programmer til STATA 11, som muliggjorde denne grafiske præsentation. Disse programmer er udarbejdet og kan findes på <http://kmd.kvfc.dk/>. Programmerne blev anvendt i Kapitel 5 til at lave figurerne for multivariat metaanalyse. Vi forventede at resultaterne af de fraktionelle polynomier vil blive lette at fortolkelige, mens graferne for de restringerede kubiske splines ville blive mere problematiske på grund af store udsving i kurverne. Vi så en tendens til dette, men vurderede at restringerede kubiske splines i vores tilfælde godt kunne fortolkes.

Del II

LINEÆRE BLANDEDE MODELLER

Blandede modeller er et udtryk for modeller, som både indeholder faste og tilfældige effekter. I dette kapitel betragtes lineære modeller med fast effekt, også kendt som fixed effect models, lineære modeller med tilfældig effekt, kendt som random effect models og blendede modeller kendt som mixed effect models. For alle disse typer af modeller antages det at responsvariablen er normalfordelt. Kapitlet tager udgangspunkt i [26].

7.1 Modeller med fast effekt

Lad $\mathbf{Y}_{n \times 1}$ betegne responsvariablen, hvor n er antallet af observationer, $\mathbf{X}_{n \times p} = (\mathbf{X}_{1(n \times 1)} \mathbf{X}_{2(n \times 1)} \dots \mathbf{X}_{p(n \times 1)})$ være en matrix bestående af forklarende variabler og $\boldsymbol{\beta} = (\beta_1 \dots \beta_p)^T$ være en vektor bestående af regressionskoefficienter, også kaldet modellens parametre. Da er den simple lineære regressionsmodel, med faste effekter, på formen:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon = \beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_p \mathbf{X}_p + \varepsilon, \quad (7.1)$$

hvor β_0 er interceptet, p er antallet af forklarende variabler, givet at hver variabel modelleres med én parameter og β_0 er interceptet og ε er den resterende fejl, også kaldet residualfejlen, som ikke forklares af de forklarende variabler $\mathbf{X}$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. At middelværdien for ε er nul, $E(\varepsilon) = 0$, betyder at den forventede værdi af en fremtidig observation y_{n+1} er $y_{n+1} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$. Modellen for responsvariablen for den j 'te person y_j kan ligeledes skrives på formen:

$$y_j \sim \mathcal{N}(\beta_0 + \beta_1 x_{1j} + \dots + \beta_p x_{pj}, \sigma^2)$$

Antallet af frie parametre $\beta_1, \dots, \beta_p$ i modellen (7.1) kaldes modellens frihedsgrader (d.f.).

Hvordan de enkelte variabler kan modelleres, kommer vi nærmere ind på i kapitel 10.

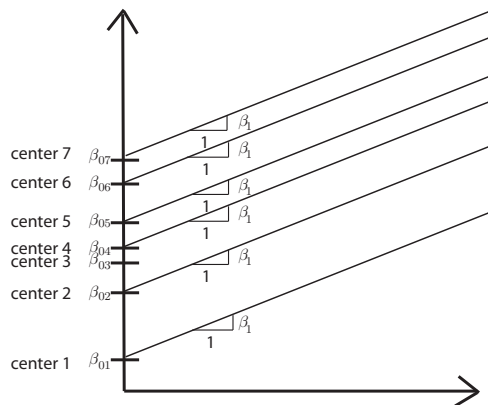
Betragtes et datasæt, hvor data er opdelt i $i = 1, \dots, N$ grupper med $j = 1, \dots, n_i$ individer i hver gruppe, kan modellen med fast effekt, hvor det antages at hver gruppe har sit eget intercept, for individ j i gruppe i , skrives på følgende form:

$$y_{ij} = \beta_{0i} + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

Hver gruppe har sit eget intercept, men samme hældning. På tilsvarende vis som interceptet kan variere for hver gruppe, kan én eller flere af parametre $\beta_1, \dots, \beta_p$, ligeledes variere for grupperne, hvilket ligeledes kaldes interaktion.

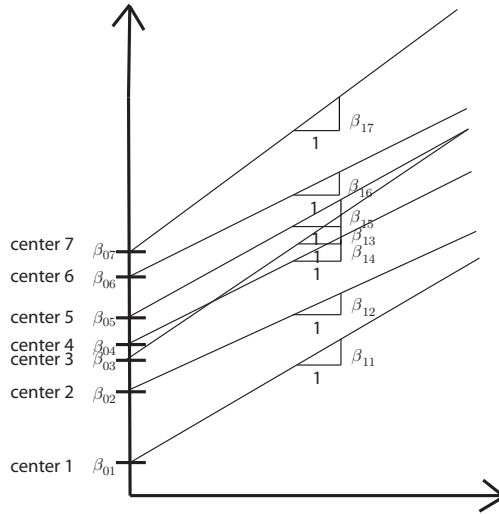
Eksempel 7.1

Betragtes et eksempel på brug af en model med fast effekt, hvor data er opdelt i syv centre og responsvariablen $\mathbf{Y}$ modelleres ved hjælp af én forklarende variabel, $\mathbf{X}$, interceptet β_{0i} variere for de syv centre men hældningen er ens for hvert af centrene, da kan $\mathbf{Y}$ skrives på formen $\mathbf{Y} = \beta_{0i} + \beta_1 \mathbf{X}$. En sådan model ses i Figur 7.1.



Figur 7.1: I figuren ses modellen $\mathbf{Y} = \beta_{0i} + \beta_1 \mathbf{X}$, hvor interceptet β_{0i} varierer for de syv centre og hældningen er ens.

Tillades ligeledes variation i hældningskoefficienten β_1 fås modellen, $\mathbf{Y} = \beta_{0i} + \beta_{1i}\mathbf{X}$. Denne model er illustreret i Figur 7.2.



Figur 7.2: I figuren ses modellen $y = \beta_{0i} + \beta_{1i}x$, hvor interceptet β_{0i} varierer for de syv centre og der tillades variation i hældningskoefficienten.

7.2 Modeller med tilfældig effekt

Modeller med tilfældig effekt kan anvendes, hvis data er opdelt i grupper. For modeller med tilfældig effekt noteres matricen, indeholdende de forklarende variabler, $\mathbf{Z}_{n \times p} = (\mathbf{Z}_{1(n \times 1)} \mathbf{Z}_{2(n \times 1)} \dots \mathbf{Z}_{p/2(n \times 1)})$ og de tilhørende modelparametre noteres $\mathbf{b} = (b_0 \ b_1 \ \dots \ b_p)^T$. Inden vi betragter modellen, hvor alle led i modellen anvender tilfældig effekt, betragtes modellen kun indeholdende et tilfældigt intercept. Modellen for deltager j i gruppe i er på formen:

$$y_{ij} = \beta_0 + b_{0i} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad b_{0i} \sim \mathcal{N}(0, \sigma_{b_0}^2). \quad (7.2)$$

hvor β_0 angiver middelniveauet for interceptet og b_{0i} er afvigelsen fra dette niveau som er trukket fra en normalfordeling. Hvis det antages at

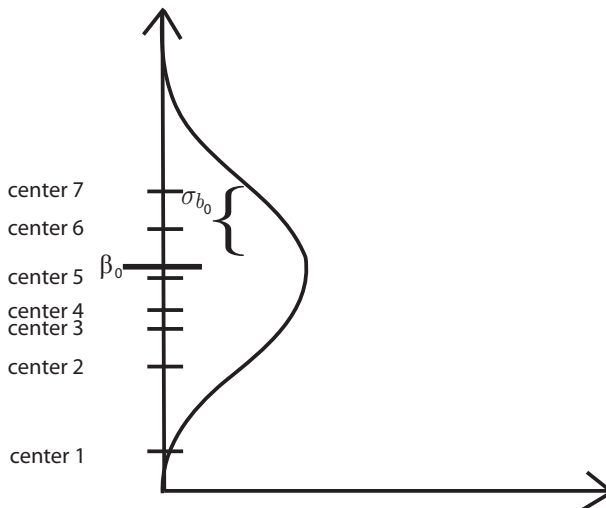
b_{0i} er uafhængig af ε_{ij} , kan modellen (7.2) ligeledes skrives som:

$$y_{ij} \sim \mathcal{N}(\beta_0, \sigma^2 + \sigma_{b_0}^2).$$

Tankegangen bag tilfældig effekt er, at de grupper man har til rådighed i sit datasæt, er en delmængde af uendelig mange normalfordelte grupper. I model (7.2) trækkes interceptet for hver gruppe fra en sådan normalfordeling, hvor man i den tilsvarende model med faste effekter som udgangspunkt har et endeligt antal grupper, som hver har et fast startniveau.

Eksempel 7.2

Betragtes igen eksemplet med de syv centre, hvor interceptet nu bestemmes ud fra en model med tilfældige effekt. Det vil sige $\mathbf{Y} = \beta_0 + b_{0i}$, $b_{0i} \sim \mathcal{N}(0, \sigma_{b_0}^2)$, $i = 1, 2, \dots, 7$. Denne model illustreres i Figur 7.3.



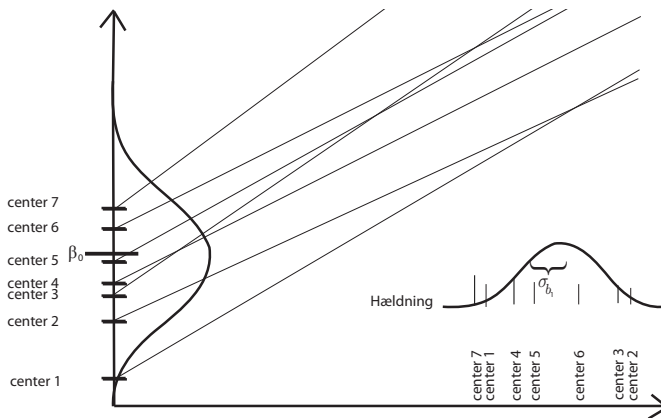
Figur 7.3: I figuren ses modellen $\mathbf{Y} = \beta_0 + b_{0i}$, $b_{0i} \sim \mathcal{N}(0, \sigma_{b_0}^2)$, hvor interceptet for de syv centre trækkes fra en normalfordeling.

Standardafvigelsen σ_{b_0} er afstanden fra middelværdien til vendetangen.]

På tilsvarende vis som ved interceptet, kan der tilføjes et tilfældigt led for hvert led i modellen.

Eksempel 7.3

Betragtes igen eksemplet med de syv centre, hvor $\mathbf{Y}$ modelleres ved hjælp af en forklarende variabel $\mathbf{Z}$ og både interceptet og hældningen er modelleres med en tilfældig effekt, fås modellen, $\mathbf{Y} = \beta_0 + b_{0i} + b_{1i}\mathbf{Z}$, $b_{0i} \sim \mathcal{N}(0, \sigma_{b_0}^2)$, $b_{1i} \sim \mathcal{N}(0, \sigma_{b_1}^2)$. Denne model illustreres som det ses, i Figur 7.4.



Figur 7.4: I figuren ses modellen $\mathbf{Y} = \beta_0 + b_{0i} + b_{1i}\mathbf{Z}$, $b_{0i} \sim \mathcal{N}(0, \sigma_{b_0}^2)$, $b_{1i} \sim \mathcal{N}(0, \sigma_{b_1}^2)$, hvor både intercept og hældning for de syv centre trækkes fra hver sin normalfordeling.

Illustrationen i nederste højre hjørne i Figur 7.4 viser fordelingen af hældningens størrelse. Det mest normale er en mellemstørrelse hældning. Disse befinder sig omkring middelværdien, mens en stor eller lille hældning er mere sjældent, hvorfor disse ses ude i halerne.

I tilfælde af indlejrede data, det vil sige data som kan opdeles i grupper i grupperne, eksempelvis byer i grupperne lande, kan man anvende mere end en tilfældig effekt til, at bestemme startniveauet for én gruppe. En model til bestemmelse af interceptet, for indlejret data, med grupperne k indlejret i grupperne i for person j , kan noteres på følgende måde:

$$y_{ijk} = \beta_0 + b_{0i} + b_{1k} + \varepsilon_{ijk},$$

hvor parameteren β_0 er det generelle niveau indenfor hver af grupperne i , $b_{0i} \sim \mathcal{N}(0, \sigma_{b_0}^2)$, $b_{1k} \sim \mathcal{N}(0, \sigma_{b_1}^2)$ og $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, hvilket betyder at alle grupperne k har samme variation, indenfor hver sin gruppe af i .

Ved at anvende modeller med tilfældige effekter, indføres antagelser om underliggende normalfordeling og ens varians. Er data ikke normalfordelt, bliver modellen med tilfældige effekter mere upræcis end den tilsvarende model med faste effekter. Fordelen ved at anvende modeller med tilfældige effekter er, at antallet af parametre, der skal estimeres, er langt færre. I tilfældet hvor kun interceptet skal estimeres, skal man i modellen med faste effekter estimere en parameter per gruppe, det vil sige N estimer. For den tilsvarende model med tilfældige effekter, skal man estimere en parameter for det faste niveau og en for variansen. Der skal altså kun estimeres to variabler, ligegyldig hvor mange grupper data inddeles i.

Antallet af parametre der anvendes kaldes også modellens kompleksitet. Hvis kompleksiteten af en model er for stor i forhold til antallet af observationer, bliver modellen ustabil. Modellens ustabilitet kan medføre, at modellen ikke kan forudsige fremtidig observationer korrekt. At modellen er for kompleks kaldes også overfitting. En simpel måde at detektere overfitting, er troværdighedsmålet.

7.2.1 Troværdighedsmålet

Troværdighedsmålet er fremkommet ved, at betragte studier, som er blevet valideret på uafhængige datasæt [14]. Disse studier har vist, at hvis antallet af parametre, p , er mindre end henholdsvis $m/10$ eller $m/20$, så er studiet, højst sandsynligt troværdigt, hvor m er den begrænsede stikprøvestørrelse, som ses i Tabel 7.1.

I dette projekt anvendes troværdighedsmålet $p < m/20$, da dette er det mindste af de to mål og dermed det der sætter de højeste krav til modellens kompleksitet.

Type af responsvariabler	Begrænset stikprøvestørrelse m
Kontinuerte	n
Binære	$\min(n_1, n_2)$
Kategoriske med k kategorier	$n - \frac{1}{n^2} \sum_{i=1}^k n_i^3$
Overlevelseshænder	antal hændelser

Tabel 7.1: I tabellen ses den begrænsede stikprøvestørrelse m for forskellige typer af responsvariabler, hvor n er den totale stikprøvestørrelse.

7.3 Blandede modeller

Blandede modeller er som tidligere nævnt en kombination af faste effekter og tilfældig effekter. Eksempelvis kan interceptet være tilfældigt, mens hældningen er fast. En blandet model kan bestå af en række af disse kombinationer.

Betragtes en blandet model for sættet af observationer $\mathbf{Y} = (y_1 \ y_2 \ \dots \ y_n)^T$ kan denne skrives på formen:

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \mathbf{b}\mathbf{Z} + \boldsymbol{\varepsilon},$$

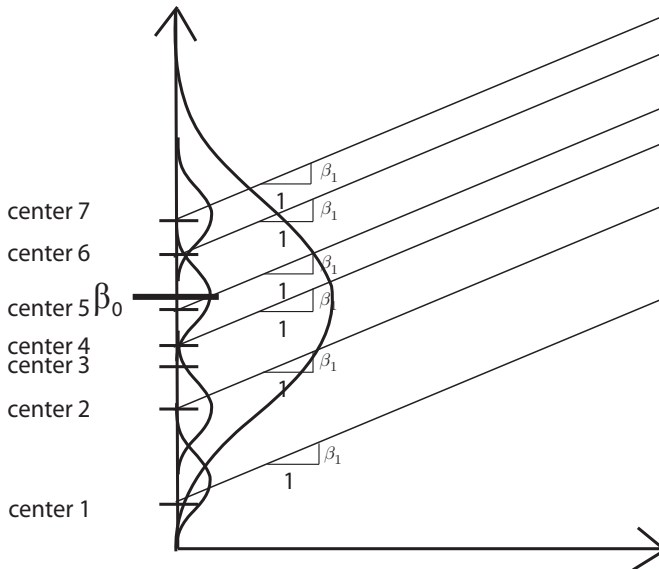
hvor $\boldsymbol{\beta} = (\beta_0 \ \beta_1 \ \dots \ \beta_{p_1})^T$ indeholder alle de faste effekter og $\mathbf{b} = (b_0 \ b_1 \ \dots \ b_{p_2})^T$ indeholder alle de tilfældige effekter og $p_1 + p_2 = p$ er antallet af modelparametre. Matricerne $\mathbf{X}_{n \times p_1}$ og $\mathbf{Z}_{n \times p_2}$ indeholder de forklarende variable for de $j = 1, \dots, n$ observationer. De tilfældige effekter $b_i \sim \mathcal{N}(0, \sigma_{b_i}^2)$ for $i = 0, \dots, p_2$ og den resterende fejl, ikke forklaret udfra de forklarende variable, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$.

Eksempel 7.4

Betragtes en blandet model for indlejret data, hvor der tages udgangspunkt i de syv centre fra tidligere, som nu er indlejret i fire lande, hvor interceptet er tilfældig og hældningen, modelleret ved en variabel $\mathbf{X}$, er fast og ens for alle centrene. Da kan responsvariablen $\mathbf{Y}_{ik}$, hvor i indekserer center og k indekserer lande, modelleres ved hjælp af modellen:

$$\mathbf{Y}_{ik} = \beta_0 + b_{0i} + b_{1k} + \beta_1 \mathbf{X}, i = 1, \dots, 7, k = 1, 2, 3, 4,$$

hvor $b_{0i} \sim \mathcal{N}(0, \sigma_{b_0}^2)$ og $b_{1k} \sim \mathcal{N}(0, \sigma_{b_1}^2)$. Modellen kan illustreres, som det ses i Figur 7.5:



Figur 7.5: I figuren ses den blandede model $\mathbf{Y}_{ik} = \beta_0 + b_{0i} + b_{1k} + \beta_1 \mathbf{X}$, $b_{0i} \sim \mathcal{N}(0, \sigma_{b_0}^2)$ og $b_{1k} \sim \mathcal{N}(0, \sigma_{b_1}^2)$, hvor interceptet for hvert af de syv centre bestemmes udfra en normalfordeling for lande og en for centre inde for hvert af landene. Alle normalfordelingerne for centrene har samme varians og hældningen er fast.

]

Alle parametrene i de ovenstående modeller er ukendte. Til at estimere parametrene, kan maksimum likelihood estimation eller restringeret maksimum likelihood estimation anvendes. Disse to metoder gennemgås i det følgende kapitel.

PARAMETER ESTIMATION

8.1 Maksimum likelihood estimation

Maksimum likelihood estimation er en almindelig anvendt metode til, at estimere ukendte parametre. Maksimum likelihood metoden giver den, eller de parametre, der er de mest sandsynlige på baggrund af det observerede. Afsnittet tager udgangspunkt i [5], [14] og [25].

Vi ønsker at estimere en vektor af ukendte parametre, ud fra et datasæt bestående af n observationer $y_1, \dots, y_n$. Da vi ikke i dette afsnit skelner mellem parametre, hørende til faste eller tilfældige effekter noteres alle parametrene β . Observationerne $y_1, \dots, y_n$ er en realisationer af stokastiske variabler for hvem sandsynligheds- eller tæthedsfunktionen for en realisation y_j noteres $f_j(y_j; \beta)$, hvor det antages at $f_j(y_j; \beta)$ er en unimodal funktion, det vil sige monotont stigende før et maksimum og monotont aftagende efter. Da er likelihoodfunktionen defineret ved:

$$\mathcal{L}(\beta) = \prod_{j=1}^n f_j(y_j; \beta).$$

Den værdi hvor $\mathcal{L}(\beta)$ antager sit maksimum, kaldes maksimum likelihood estimatet for β og noteres $\hat{\beta}$.

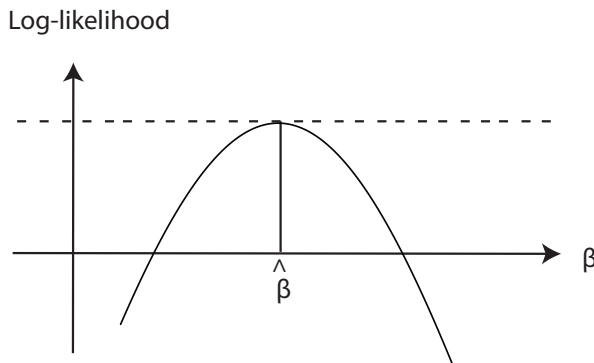
For at finde maksimum likelihood estimatet betragtes alle mulige udfald og maksimum af funktionen findes. Ofte anvendes log-likelihoodfunktionen i stedet for likelihoodfunktionen, da denne er analytisk nemmere at bearbejde. Log-likelihoodfunktionen er defineret ved:

$$\ell(\beta) = \log \mathcal{L}(\beta) = \sum_{j=1}^n \log f_j(y_j; \beta),$$

I stedet for at maksimere likelihoodfunktionen, maksimeres log-likelihoodfunktionen, idet disse to funktioner har samme maksimums værdi. Fremgangsmåden for maksimum likelihood estimation er som følger:

- I Find likelihoodfunktionen, $\mathcal{L}(\beta)$, og log-likelihoodfunktionen, $\ell(\beta)$.
- II Differentier $\ell(\beta)$ partielt med hensyn til β og sæt lig med nul. Den differentierede log-likelihoodfunktionen findes ved $\ell'(\beta) = (\partial/\partial\beta)\ell(\beta)$ og noteres $S(\beta)$. Den differentierede log-likelihoodfunktionen kaldes også scorefunktionen.
- III Løs $S(\beta) = 0$ for β og kald løsningen $\hat{\beta}$. Løsningen $\hat{\beta}$ er maksimum likelihood estimatet af β .

I Figur 8.1 ses en grafisk illustration af, hvordan maksimum likelihood estimatet bestemmes for én ukendt parameter β i tilfælde af ét globalt maksimum, hvilket følger af antagelsen af, at funktionen er unimodal.



Figur 8.1: I figuren ses maksimum likelihood estimatoren $\hat{\beta}$.

Da det antages at tæthedsfunktionen er unimodal, vil der altid være ét maksimum, som det ses i Figur 8.1. Havde denne antagelse ikke været opfyldt således at der kunne forekomme flere maksima,

likelihoodfunktionen kunne være konstant eller tangerede en maksimumsværdi, men aldrig nåede en, da ville maksimum likelihood estimatet $\hat{\beta}$ ikke kunne bestemmes. Har man en forhåndsviden om hvor maksimum bør ligge, kan man, i visse tilfælde, finde maksimum likelihoodestimatet for en likelihoodfunktion med flere maksima.

I nogle tilfælde er det ikke muligt at løse $S(\beta) = 0$ eksplicit. Derfor må numeriske metoder tages i brug, en metode til dette er Newton-Raphson metoden, som er baseret på at approksimere $S(\beta)$ med en ret linie, indenfor et lille område. Den lineære approksimation for start estimatet, er givet ved:

$$S(\beta) = S(\beta^0) - I(\beta^0)(\beta - \beta^0),$$

hvor β^0 er start estimatet og $I(\cdot)$ er den anden afledte af log-likelihoodfunktionen som kaldes Fishers information.

Den approksimerede score funktion sættes lig med nul vektoren og løses med hensyn til β ,

$$\beta = \beta^0 + I^{-1}(\beta^0)S(\beta^0),$$

og β anvendes som det nye estimat β^1 . Løsningen for den k 'te iteration bliver da:

$$\beta^{k+1} = \beta^k + I^{-1}(\beta^k)S(\beta^k).$$

Newton-Raphson iterationen fortsætter indtil ændringerne for $-2 \log$ -likelihood bliver mindre end en tolerance tærskel.

Likelihoodfunktionen for modellen med faste effekter (7.1) er givet ved:

$$\mathcal{L}(\beta) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y_j - (\beta_0 + \beta_1 x_{1j} + \dots + \beta_p x_{pj}))^2}{-2\sigma^2}\right),$$

hvor $\beta = (\beta_0 \ \beta_1 \ \dots \ \beta_p)^T$.

Likelihoodfunktionen for modellen med tilfældige effekter (7.2) er givet ved:

$$\mathcal{L}(\beta) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_{b_0}^2)}} \exp\left(\frac{(y_j - \alpha_0)^2}{-2(\sigma^2 + \sigma_{b_0}^2)}\right).$$

Under generelle betingelser gælder derfor maksimum likelihood estimation, at metoden er konsistent, hvilket vil sige at når stikprøvestørrelsen stiger, kommer maksimum likelihoodestimatet tættere på den sande værdi. Det gælder ligeledes for store stikprøvestørrelser, at likelihood-funktionen er approksimativt normalfordelt med middelværdi i den sande parameter værdi. At likelihood funktionen er approksimativt normalfordelt, kan avendes til at bestemme konfidensintervallet for maksimum likelihood estimatet. Konfidensintervallet for en model tilføjes for, at indikere præcisionen af modellen. Dette er nødvendigt idet en model aldrig er korrekt. Den kan passe mere eller mindre godt med den sande sammenhæng, men kan aldrig bestemme den sande sammenhæng. Hvor godt modellen stemmer overens med virkeligheden, kaldes modellens præcision. Forskellige mål for præcisionen kan anvendes til, at vurdere hvorvidt en model er bedre end en anden.

8.1.1 Akaike's informations kriterium

Hvis flere forskellige modeller er mulige til at modellere samme respons, kan Akaike's informations kriterium anvendes til at bestemme hvilken model, der skal benyttes.

Akaike's informations kriterium (AIC) kan benyttes til at vurdere hvilken model, der giver det bedste forhold mellem modellens evne til at forudsige fremtidige observationer og dens kompleksitet. Modellens evne til at forudsige fremtidige observationer, henføres også til som modellens evne til, at prædiktere. Akaike's informations kriterium er defineret ud fra devians differensen mellem nul-modellen,

$\beta_0 = \dots = \beta_p = 0$, og modellen hvor parametervektoren β estimeres med maksimum likelihood estimatet, $\hat{\beta}$, samt modellens frihedsgrader.

Deviansen defineres som:

$$\text{Devians} = D = -2\ell,$$

Devians differensen mellem to modeller i og j er da defineret som:

$$D_i - D_j = -2\ell_i - (-2\ell_j), \quad (8.1)$$

Devians differensen mellem nul-modellen og modellen hvor β estimeres med maksimum likelihood estimatet, $\hat{\beta}$, kaldes likelihoodratioen og noteres LR . Deviansen for nul-modellen notered $D_{\mathbf{0}} = -2\ell(\mathbf{0})$. Likelihoodratioen er da givet ved:

$$LR = D_{\mathbf{0}} - D_{\hat{\beta}} = -2\ell(\mathbf{0}) + 2\ell(\hat{\beta}) = -2 \log \left(\frac{\mathcal{L}(\mathbf{0})}{\mathcal{L}(\hat{\beta})} \right),$$

Likelihoodratioen er approksimativt χ^2 fordelt med $p-1$ frihedsgrader, hvor p er antallet af model parametre.

Akaike's informations kriterium er da defineret ved:

$$AIC = LR - 2 \times d.f$$

Den model med den højeste AIC, giver det bedste forhold mellem modellen kompleksitet og dens præcision.

8.2 Restringeret maksimum likelihood

Ved modeller med tilfældig effekter samt blandede modeller, skal varians parametre estimeres. Ved anvendelse af maksimum likelihood estimation tages der højde for tabet af frihedsgrader ved estimation af

de faste effekter, hvorfor estimatet for varians komponenterne bliver for lille. En metode til at undgå bias ved estimation af varians- eller kovariansparametre, er restringeret maksimum likelihood estimation

Restringeret maksimum likelihood (REML) er en tilgang til maksimum likelihood estimation, som ikke anvender det fulde sæt af observationer. I stedet anvendes en likelihoodfunktion udregnet fra et transformeret sæt af observationer således, at de parametre som er uvæsentlige i forhold til variansen, ikke har nogen effekt på variansen, det vil sige de faste parametre. Afsnittet tager udgangspunkt i [15] og [38].

Til at illustrere restringeret maksimum likelihood estimation, betragtes en bandet model for sættet af observationer

$$\mathbf{Y} = (y_1 \ y_2 \ \dots \ y_n)^T:$$

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \mathbf{b}\mathbf{Z} + \boldsymbol{\varepsilon},$$

hvor $\boldsymbol{\beta} = (\beta_0 \ \beta_1 \ \dots \ \beta_{p_1})^T$ indeholder alle de faste effekter og $\mathbf{b} = (b_0 \ b_1 \ \dots \ b_{p_2})^T$ indeholder alle de tilfældige effekter. Matricerne $\mathbf{X}_{n \times p_1}$ og $\mathbf{Z}_{n \times p_2}$ indeholder de forklarende variable for de $j = 1, \dots, n$ observationer. De tilfældige effekter $b_i \sim \mathcal{N}(0, \sigma_{b_i}^2)$ for $i = 0, \dots, p_2$ og den resterende fejl, ikke forklaret udfra de forklarende variabler, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$.

Sættet af observationer $\mathbf{Y}$ transformeres ved, at finde en matrix $\mathbf{A}_{(n \times (n-p_1))}$, som er ortogonal på $\mathbf{X}_{n \times p_1}$, således at $\mathbf{A}^T \mathbf{X} = 0$. Ved at multiplicere sættet af observationer $\mathbf{Y}$ med $\mathbf{A}$, fjernes effekten af de faste parametre. Det transformerede sæt af observationer noteres $\tilde{\mathbf{Y}}$ og er da givet ved:

$$\tilde{\mathbf{Y}} = \mathbf{A}^T \mathbf{Y}.$$

Efter transformationen er fremgangsmåden for estimation af parametre ved restringeret maksimum likelihood estimation, den samme som ved maksimum likelihood estimation. Den restringerede likelihoodfunktion og log-likelihoodfunktion er givet ved:

$$\mathcal{L}_r(\boldsymbol{\beta}) = \prod_{j=1}^n f_j(\tilde{y}_j; \boldsymbol{\beta})$$

og

$$\ell_r(\boldsymbol{\beta}) = \log \mathcal{L}_r(\boldsymbol{\beta}) = \sum_{j=1}^n \log \mathcal{L}_{r_j}(\boldsymbol{\beta}).$$

Et problem med restringeret maksimum likelihood estimation er ,at Akaike's informations kriterium ikke kan anvendes. Dette skyldes at ved restringeret maksimum likelihood estimation anvendes forskellige sæt af observationer for de forskellige modeller, hvormed deres middelværdistrukturer bliver forskellige. Derved bliver deres log-likelihoodfunktioner ikke sammenlignelige. Derfor anvendes maksimum likelihood estimation til modelselektion, hvorefter de endelige modelparametre for blandede modeller og modeller med tilfældig effekt, estimeres ved hjælp af restringeret maksimum likelihood estimation.

I denne rapport betragtes der, udover blandede modeller, ligeledes en anden metode, til at opnå et samlet estimat for data inddelt i grupper. Denne metode kaldes metaanalyse. Metaanalyse kan først anvendes efter modelparametrene, for hver af grupperne, er blevet estimeret. Metoden metaanalyse gennemgås i det næste kapitel. I forbindelse med metaanalyse, vil regressionskoefficienterne, hørende til eksponeringen, ligeledes blive refereret til som den modellerede effekt.

METAANALYSE

Metaanalyse er en metode til, at kombinere associationsmål fra flere studier, til et samlet associationsmål. Metaanalyse anvendes, når flere studier betragter den samme association. Dette kapitel vil gennemgå univariat metaanalyse med fast effekt, også kendt som fixed metaanalysis, univariat metaanalyse med tilfældig effekt, også kendt som random metaanalysis, multivariat metaanalyse med fast effekt samt multivariat metaanalyse med tilfældig effekt. Kapitlet tager udgangspunkt i [21], [7], [30] og [35].

Metaanalyse er blevet indført for at kunne få et systematisk overblik over alle offentliggjorte resultater indenfor et bestemt område. Ved at kombinere resultaterne fra alle kendte studier fremfor, at betragte ét studie, opnås et bredere billede af, hvordan responsvariablen er associeret med interessevariablen. For at få det mest korrekte billede af effekten, kræver det at alle offentliggjorde studier og ikke kun af dem, der er mest interessante, relevante, eller let tilgængelige indsamles og analyseres. Den ofte store mængde indsamlet information indeholder ofte ikke-konsistente konklusioner og man bør derfor være kritisk når man kombinerer resultaterne af de indsamlede studier.

Kombinationen af resultaterne fra de mange studier er ikke et simpelt gennemsnit, men et vægtet gennemsnit. Alt efter om metaanalysen er univariat eller multivariat og har fast eller tilfældig effekt, bestemmes vægtene forskelligt.

9.1 Univariat metaanalyse med fast effekt

Målet af interesse er den sande værdi β . Til at estimere β betragtes estimaterne fra N studier, $\hat{\beta}_i$ for $i = 1, \dots, N$. Det endelige estimat af β findes som et vægtet gennemsnit af $\hat{\beta}_i$ -værdierne. Det vægtede gennemsnit noteres $\widehat{\beta}_{\text{fa}}$, for univariat metaanalyse med fast effekt, og findes ved:

$$\bar{\beta}_{\text{fa}} = \frac{\sum_{i=1}^N w_i \hat{\beta}_i}{\sum_{i=1}^N w_i}, \quad (9.1)$$

hvor nævneren anvendes til at normalisere vægten w_i .

Ligeegyldig hvilken metaanalyse-metode man vælger, vil det vægtede gennemsnit kunne skrives på formen (9.1).

For univariat metaanalyse med fast effekt bestemmes w_i ud fra variansen for parameter estimatet $\hat{\beta}_i$ indenfor hvert studie, $\sigma_{w_i}^2(\hat{\beta})$.

$$w_i = \frac{1}{\sigma_{w_i}^2(\hat{\beta})}.$$

Hvis et studie har en stor varians får den derved en lille vægt ved bestemmelse af det endelige resultat af metaanalysen.

For at kunne bestemme det vægtede gennemsnit $\bar{\beta}_{\text{fa}}$, antages det at $\hat{\beta}_i$ og $\sigma_{w_i}^2(\hat{\beta})$ er kendte.

Under antagelse af, at parameterestimerne er normalfordelte og antallet af studier $N \rightarrow \infty$, kan konfidensintervallet for β bestemmes som:

$$95\%CI(\beta) = \bar{\beta}_{\text{fa}} \pm 1,96 \frac{1}{\sqrt{\sum_{i=1}^N w_i}}.$$

Eksempel 9.1

Betragt et eksempel, hvor syv studier studerer sammenhængen mellem alder og ændringer i livvidde/højde. Først indsamles information om målet for denne sammenhæng, samt den tilhørende varians. Disse ses i Tabel 9.1.

Center	$\hat{\beta}$	varians $v(\hat{\beta})$	vægt w
Firenze	-0,279	0,003	293,78
Cambridge	-0,185	0,001	967,71
Amsterdam	-0,590	0,006	163,78
Doetinchem	-0,229	0,004	250,92
Maastricht	-0,719	0,006	173,87
Århus	-0,050	0,008	127,59
København	0,050	0,004	264,43
Sum			2.081,47

Tabel 9.1: I tabellen ses estimer, varians samt vægt for de syv centre, Firenze (Italien), Cambridge (England), Amsterdam (Holland), Doetinchem (Holland), Maastricht (Holland), Århus (Danmark), København (Danmark).

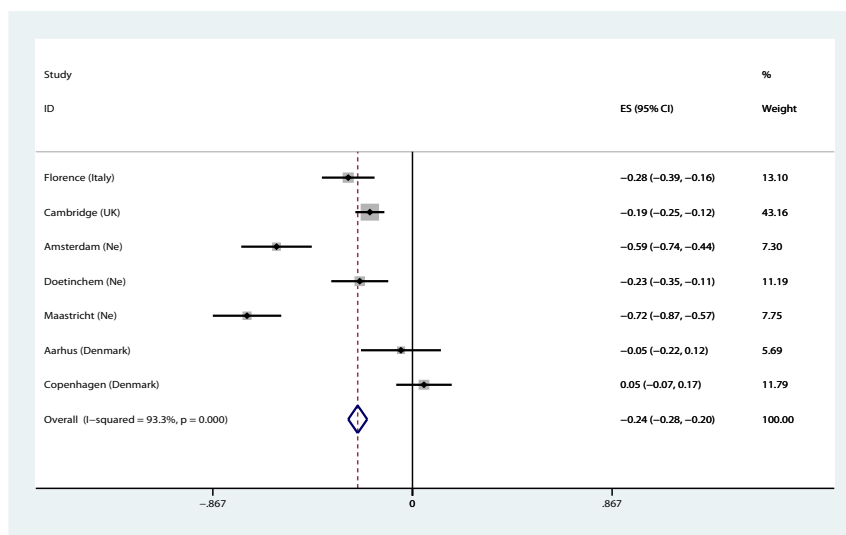
Derefter udregnes det vægtede gennemsnit (9.1) og det tilhørende 95% konfidensinterval:

$$\begin{aligned}\bar{\hat{\beta}}_{\text{fa}} &= \frac{-0,28 \cdot 293,78 - 0,19 \cdot 967,71 - 0,59 \cdot 163,78 - 0,23 \cdot 250,92}{2.081,47} \\ &\quad - \frac{-0,72 \cdot 173,87 - 0,05 \cdot 127,59 + 0,05 \cdot 264,43}{2.081,47} \\ &= -0,24\end{aligned}$$

$$95\%CI(\beta) = -0,24 \pm 1,96 \frac{1}{\sqrt{2.081,47}} = (-0,28; -0,20).$$

En metode til grafisk, at gøre den univariate metaanalyse overskuelig er, ved hjælp af et forest plot. Et forest plot for metaanalysen for de syv studier i dette eksempel, ses i Figur 9.1.

Hver af studierne er illustreret ved hjælp af et punkt for punktestimatet, en boks der fortæller hvor stor variation der er i studiet, samt en vandret streg, der illustrerer konfidensintervallet for studiet. Løsriver konfidensintervallet sig fra den lodrette streg, som går gennem nul, er effekten af interessevariablen signifikant. I bunden af plottet ses det samlede estimat med tilhørende konfidensinterval, illustreret ved



Figur 9.1: I figuren ses en grafisk illustration af en univariat metaanalyse med fast effekt.

en rombe. I højre side er de præcise værdier for alle punktestimater og konfidensinterval opskrevet. Vægtene i figuren er normaliserede til at summere til 100. Gøres det samme for vægtene i Tabel 9.1 opnås samme vægte.

I venstre side af plottet er studiernes navne skrevet op. Udfra overall, ses p-værdien for heterogenitets testet samt en I-squared værdi. I-squared er et andet test for heterogenitet, som fortæller hvor mange procent heterogenitet der er mellem studierne.]

Ved at betragte forest plottet, kan man få en ide om, om effekten er homogen. For at fastslå om studierne er homogene, kan et test for heterogenitet anvendes. Hvis det viser sig at studierne er heterogene, argumenterer [7] for at metaanalyse med tilfældig effekt er mere virkelighedstro end metaanalyse med fast effekt.

Teststørrelsen for heterogenitetstestet udregnes som følger:

$$Q = \sum_{i=1}^N (w_i(\hat{\beta}_i - \bar{\hat{\beta}}_{\text{fa}})^2),$$

Under nulhypotesen:

$$H_0: \hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_N$$

er Q approksimativt χ^2 fordelt med $N - 1$ frihedsgrader såfremt at $\hat{\beta}_i \sim \mathcal{N}(\beta, \sigma^2)$.

I-squared findes ved:

$$I^2 = \frac{Q - d.f.}{Q} 100\%,$$

hvor $d.f. = N - 1$.

Ifølge [35] er antagelse af homogenitet altid tvivlsom, selvom homogeniteten kan være svært at modbevise for metaanalyser med få studier. De mener, at der kan være heterogenitet til stede, selv hvis test for heterogenitet ikke er signifikant og heterogenitet altid bør være en del af analysen. Deres argument er, at heterogenitet forventes at være til stede, da de enkelte undersøgelser ikke er identiske med hensyn til studiepopulation samt andre faktorer, som kan forårsage forskelle mellem undersøgelserne.

Vi mener at man bør overveje hvorvidt ens data kan betragtes som et stort eller flere små studier inden man vælger analysestrategi. Heterogenitetstestet kan være en del af denne vurdering, men kan ikke stå alene.

9.2 Univariat metaanalyse med tilfældig effekt

For metaanalyse med tilfældig effekt antager man, at effekten samlet set er ens og at det skyldes tilfældig variation, at de enkelte estimater

ikke er magen til estimatet $\widehat{\beta}_{\text{tf}}$, som er estimatet af den sande effekt fundet ved det vægtede gennemsnit for metaanalyse med tilfældig effekt.

Betragt igen N studier, hvor $\widehat{\beta}_i$ er estimatet for målet af interesse β_i , hvor $i = 1, \dots, N$. Hvis det antages at $\widehat{\beta}_i$ 'erne og β_i 'erne er normalfordelte, hvor β_i er den sande underlæggende værdi i hvert studie, kan $\widehat{\beta}_i$ udtrykkes ved:

$$\widehat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma_{w_i}^2(\widehat{\beta})),$$

hvor

$$\beta_i \sim \mathcal{N}(\beta, \sigma_b^2(\widehat{\beta})).$$

Hvis det antages at variansen i hvert studie $\sigma_{w_i}^2(\widehat{\beta})$ og variansen mellem studierne $\sigma_b^2(\widehat{\beta})$ er uafhængige, kan $\widehat{\beta}_i$ ligeledes udtrykkes ved:

$$\widehat{\beta}_i \sim \mathcal{N}(\beta, \sigma_{w_i}^2(\widehat{\beta}) + \sigma_b^2(\widehat{\beta})).$$

Det antages altså at hvert studie har en sand værdi β_i og variansen indenfor hvert studie kan forklares ud fra variansen $\sigma_{w_i}^2(\widehat{\beta})$. Den sande værdi β_i er trukket fra en normalfordeling med middelværdi β og variationen mellem studierne kan forklares ud fra variansen $\sigma_b^2(\widehat{\beta})$.

9.2.1 Estimation af de ukendte parametre

Hvis det antages at $\widehat{\beta}_i$ og $\sigma_{w_i}^2(\widehat{\beta})$ er kendte, skal de ukendte parametre $\sigma_b^2(\widehat{\beta})$ og β estimeres. Estimatet af β svarer til det vægtede gennemsnit $\widehat{\beta}_{\text{tf}}$.

Parametrene estimeres ved hjælp af maksimum likelihood estimation eller restringeret maksimum likelihood estimation. Da det antages at $\widehat{\beta}_i$ er normalfordelt er likelihood funktionen for $\widehat{\beta}$ givet ved:

$$\mathcal{L}(\widehat{\beta}) = \prod_{i=1}^N \frac{1}{2\pi(\sigma_{w_i}^2(\widehat{\beta}) + \sigma_b^2(\widehat{\beta}))} \exp \left(-\frac{\widehat{\beta}_i - \beta}{2(\sigma_{w_i}^2(\widehat{\beta}) + \sigma_b^2(\widehat{\beta}))} \right)^2,$$

og log-likelihoodfunktionen er givet ved:

$$\ell(\hat{\beta}) = \sum_{i=1}^N \log\left(\frac{1}{2\pi}\right) + \log\left(\frac{1}{\sigma_{w_i}^2(\hat{\beta}) + \sigma_b^2(\hat{\beta})}\right) + \left(\frac{\hat{\beta}_i - \beta}{-2(\sigma_{w_i}^2(\hat{\beta}) + \sigma_b^2(\hat{\beta}))}\right)^2. \quad (9.2)$$

Ved at maksimere (9.2) fås $\bar{\hat{\beta}}$ udtrykt ved den estimerede varians parameter $\hat{\sigma}_b^2(\hat{\beta})$:

$$\bar{\hat{\beta}}_{\text{tf}} = \frac{\sum_{i=1}^N \frac{\hat{\beta}_i}{\sum_{i=1}^N \sigma_{w_i}^2(\hat{\beta}) + \hat{\sigma}_b^2(\hat{\beta})}}{\frac{1}{\sigma_{w_i}^2(\hat{\beta}) + \hat{\sigma}_b^2(\hat{\beta})}} = \frac{\sum_{i=1}^N w_i \hat{\beta}_i}{\sum_{i=1}^N w_i}, \quad (9.3)$$

hvor $w_i = \frac{1}{\sigma_{w_i}^2(\hat{\beta}) + \hat{\sigma}_b^2(\hat{\beta})}$.

Da log-likelihoodfunktionen (9.2) ikke kan løses implicit, kaldes den en pseudolikelihoodfunktion og parameterestimererne kan findes ved hjælp af Newton-Raphson metoden, beskrevet i afsnit 8.1. Da varians parameteren $\sigma_b^2(\hat{\beta})$ skal bestemmes, kan det være en fordel at anvende den restringerede likelihoodfunktion til at estimere parametrene for at undgå at estimatet bliver for lille. Den restringerede log-likelihoodfunktion for $\hat{\beta}$ er givet ved:

$$\ell_r(\hat{\beta}) = \ell(\hat{\beta}) - \log\left(\sum_{i=1}^N \frac{1}{\sigma_{w_i}^2(\hat{\beta}) + \sigma_b^2(\hat{\beta})}\right) - \log\left(\frac{1}{2\pi}\right).$$

Konfidensintervallet for β bestemmes ved:

$$95\%CI(\beta) = \bar{\hat{\beta}}_{\text{tf}} \pm 1,96 \frac{1}{\sqrt{\sum_{i=1}^N w_i}} = \bar{\hat{\beta}}_{\text{tf}} \pm 1,96 \frac{1}{\sqrt{\sum_{i=1}^N \frac{1}{\sigma_{w_i}^2(\hat{\beta}) + \hat{\sigma}_b^2(\hat{\beta})}}}.$$

Eksempel 9.2

Testes for heterogenitet i de syv studier der betragter sammenhængen mellem alder og ændringer i livvidde/højde fra Eksempel 9.1, fås at:

$$\begin{aligned}
Q &= 293,78(-0,279 + 0,24)^2 + 967,71(-0,185 + 0,24)^2 + 163,78(-0,590 + 0,24)^2 \\
&\quad + 250,92(-0,229 + 0,24)^2 + 173,87(-0,719 + 0,24)^2 + 127,59(-0,050 + 0,24)^2 \\
&\quad + 264,43(0,050 + 0,24)^2 \\
&= 90,21
\end{aligned}$$

og

$$I^2 = \frac{90,21 - 6}{90,21} 100\% = 93\%$$

ved at slå op i tabellen for χ^2_6 ses at p-værdien $p < 0,001$. Det vil sige at nulhypotesen:

$$H_0: \text{Studierne er homogene: } \hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_7,$$

afvises. Derfor anvendes metaanalyse med tilfældig effekt. Værdierne for $\hat{\beta}_i$ og $\sigma^2_{w_i}(\hat{\beta})$ er de samme som i eksemplet 9.1, disse ses i Tabel 9.1. Variansen mellem studierne $\sigma^2_b(\hat{\beta})$ estimeres ved hjælp af restringeret likelihood estimation til:

$$\hat{\sigma}^2_b(\hat{\beta}) = 0,72.$$

Dermed bliver vægtene for metaanalysen, med tilfældig effekt, som det ses i Tabel 9.2.

Center	vægt w
Firenze	1,40
Cambridge	1,56
Amsterdam	1,33
Doetinchem	1,38
Maastricht	1,34
Århus	1,29
København	1,39
Sum	9,69

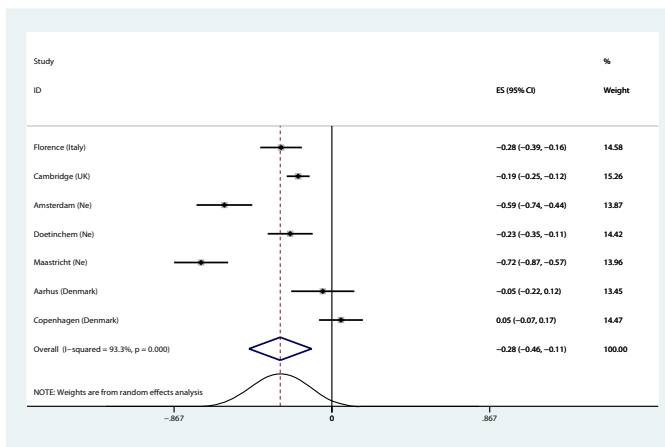
Tabel 9.2: I tabellen vægtene for de syv centre, Firenze (Italien), Cambridge (England), Amsterdam (Holland), Doetinchem (Holland), Maastricht (Holland), Århus (Danmark), København (Danmark).

Derefter udregnes det vægtede gennemsnit (9.3) og det tilhørende 95% konfidensinterval:

$$\begin{aligned}\bar{\beta}_{\text{tf}} &= \frac{-0,28 \cdot 1,40 - 0,19 \cdot 1,56 - 0,59 \cdot 1,33 - 0,23 \cdot 1,38}{9,69} \\ &= \frac{-0,72 \cdot 1,34 - 0,05 \cdot 1,29 + 0,05 \cdot 1,39}{9,69} \\ &= -0,28\end{aligned}$$

$$\begin{aligned}95\%CI(\bar{\beta}_{\text{tf}}) &= -0,28 \pm 1,96 \frac{1}{\sqrt{9,69}} \\ &= (-0,46; -0,11).\end{aligned}$$

Den univariate metaanalyse med tilfældig effekt kan på tilsvarende vis som den univariate metaanalyse med fast effekt, illustreres ved hjælp af et forest plot. Et forest plot for metaanalysen for de syv studier, ses i Figur 9.2. På figuren er normalfordelingen $\mathcal{N}(\beta, \sigma_{w_i}^2 + \sigma_b^2) = \mathcal{N}(-0,28; 0,09)$ indtegnet, hvilket er den fordeling er den de syv studier er trukket fra.



Figur 9.2: I figuren ses en grafisk illustration af en univariat metaanalyse med tilfældig effekt.

Hvis vi sammenligner med resultaterne fra eksempel 9.1, hvor $\widehat{\beta}_{\text{fa}} = -0,24 (-0,28; -0,20)$, ser vi at punktestimaterne næsten er det samme, men at konfidensintervallerne bliver bredere ved brug af metaanalyse med tilfældig effekt.]

9.3 Multivariat metaanalyse med fast effekt

Ved multivariat metaanalyse med fast effekt, betragtes flere sammenhængende variable af interesse $\widehat{\beta}_{ij}$, hvor i er antallet af studier $i = 1, \dots, N$ og j er antallet af variabler man ønsker at betragte $j = 1, \dots, m$. Hvis variablerne betragtes individuelt udtrykkes informationen om deres sammenhæng ikke. For at udnytte denne information, kan man lave en multivariat metaanalyse. Ved multivariat metaanalyse med fast effekt anvendes en vektor $\widehat{\beta}_i$, af længde m , indeholdende estimerne $\widehat{\beta}_{ij}$ af variablerne af interesse β_{ij} , for hvert studie, samt den tilhørende kovariansmatricen $\Sigma_{i(m \times m)}$ til bestemmelse resultatvektoren $\widehat{\beta}_{\text{fa}}$.

Vægt matricen $\mathbf{W}_i$ bestemmes ud fra kovariansmatricen Σ_i til:

$$\mathbf{W}_i = \Sigma_i^{-1}.$$

Hvis det antages at $\widehat{\beta}_i$ og Σ_i er kendte kan de vægtede gennemsnit findes ved:

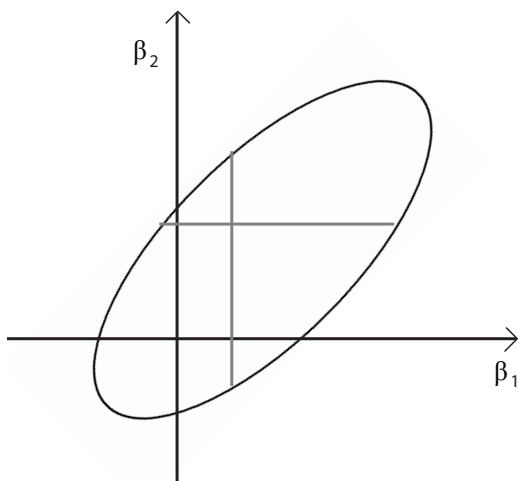
$$\widehat{\beta}_{\text{fa}} = \left(\sum_{i=1}^N \Sigma_i^{-1} \right)^{-1} \left(\sum_{i=1}^N \Sigma_i^{-1} \widehat{\beta}_i \right)$$

og kovariansmatricen for β findes ved:

$$\text{cov}(\beta) = \widehat{\Sigma} = \left(\sum_{i=1}^N \Sigma_i^{-1} \right)^{-1}$$

Hvilket kan anvendes til at bestemme en konfidensregion for β . En konfidensregion er et flere-dimensionalt konfidensinterval. For to

dimensioner kunne en konfidensregion se ud som i Figur 9.3. Der er 95 procent sandsynlighed for at den sande $\beta = (\beta_1 \beta_2)$ ligger indenfor regionen. Hvis en af parametrene holdes fast svarer de går markeringer i Figuren til konfidensbåndene for den ene variabel justeret for den anden.



Figur 9.3: I figuren ses konfidensregionen for β_1 og β_2 .

Værktøjer til analyse af det univariate tilfælde af metaanalyse er inkorporeret i mange statistik-softwareprogrammer, herunder STATA 11, som anvendes til analyser i dette projekt. Når det kommer til værktøjer til multivariat metaanalyse, er det mere begrænset, hvad der er til rådighed. Til at lave multivariat metaanalyse i STATA 11, kan kommandoerne *mvmeta_make* og *mvmeta*, skrevet af Ian White, installeres [36]. Kommandoen er ikke standard i de nuværende STATA 11 udgaver. Ian Whites kommando er anvendt til at lave multivariat metaanalyse i dette projekt.

Vi har introduceret multivariat metaanalyse, idet det bliver nødvendigt for at kunne lave metaanalyse, for ikke-lineær modellering af kontinuerte variabler. Da vi endnu ikke har gennemgået de ikke-lineære modelleringsmetoder, har vi medtaget et eksempel, for to tilfældigt

udvalgte variabler. Fortolkningen af den multivariate metaanalyse, i det følgende eksempel, svarer til univariate metaanalyser, hvor der er blevet indbyrdes justeret for de implicerede kovariater. Fremgangsmåden for denne multivariate metaanalyse er den samme som når vi introducere ikke-lineær modellering.

Eksempel 9.3

Betragtes igen de syv centre, hvor vi undersøger sammenhængen mellem alder og ændringer i livvidde/høje, men nu ønsker vi samtidig at betragte indtag af fed fisk i forhold til ændringer i livvidde/høje. Vi undersøger hvad der sker ved at øge indtaget af fed fisk med 10 gram om dagen. Vi vurderer at der er en sammenhæng mellem alder og hvor meget fed fisk man indtager og ønsker at tage denne sammenhæng med i betragtning i metaanalyserne. Derfor laves multivariat metaanalyse. I Tabel 9.3 ses punktestimaterne samt den tilhørende varians for alder og fed fisk samt kovariansen mellem alder og fed fisk.

Center	$\hat{\beta}_{\text{alder}}$	$\hat{\beta}_{\text{fed fisk}}$	$v(\hat{\beta})_{\text{alder}}$	$v(\hat{\beta})_{\text{fed fisk}}$	$cov_{\text{alder/fed fisk}}$
Firenze	-0,28	0,27	0,003	0,20	0,001
Cambridge	-0,18	0,18	0,001	0,04	0,0003
Amsterdam	-0,59	-1,74	0,01	2,93	0,01
Doetinchem	-0,23	0,98	0,004	3,37	-0,01
Maastricht	-0,72	-2,82	0,01	5,96	0,01
Århus	0,01	-1,67	0,01	0,07	-0,003
København	0,10	-1,40	0,004	0,03	-0,001

Tabel 9.3: I tabellen ses estimer, varians samt kovarians for alder og indtag af fed fisk for de syv centre, Firenze (Italien), Cambridge (England), Amsterdam (Holland), Doetinchem (Holland), Maastricht (Holland), Århus (Danmark), København (Danmark).

Det vægtede gennemsnit er fundet ved hjælp af STATA 11 til:

$$\widehat{\bar{\beta}}_{\text{fa}} = (-0,24 - 0,72)^T,$$

hvor første estimat hører til alder og andet estimat hører til 10 gram

øget indtag af fed fisk.

Konfidensintervallet for β_{alder} , for fastholdt indtag af fed fisk, bliver da:

$$95\%CI(\widehat{\beta}_{\text{fa}_{\text{alder}}}) = (-0,28; -0,20)$$

og konfidensintervallet for $\beta_{\text{fed fisk}}$, for fastholdt alder bliver:

$$95\%CI(\widehat{\beta}_{\text{fa}_{\text{fed fisk}}}) = (-0,94; 0,50)$$

]

9.4 Multivariat metaanalyse med tilfældig effekt

Som ved multivariat metaanalyse med fast effekt, betragtes flere sammenhængende variabler af interesse. Det antages, at hver af variablerne har én sand værdi, trukket fra en multivariat normalfordeling og afvigelserne fra denne værdi skyldes tilfældig variation.

Betragt igen N studier, hvor $\widehat{\beta}_{ij}$ er estimeret af målene af interesse β_{ij} , hvor $i = 1, \dots, N$ og $j = 1, \dots, m$. Hvis det antages at vektoren $\widehat{\beta}_i$ er multivariat normalfordelt, kan $\widehat{\beta}_i$ udtrykkes ved:

$$\widehat{\beta}_i \sim \mathcal{N}(\beta, \Sigma_i + \Omega).$$

hvor β er en vektor, af længde m , bestående af de sande underlæggende værdier, Σ_i er en $m \times m$ kovariansmatrix som indeholder variansen og kovariansen indenfor hvert studie og Ω er en $m \times m$ kovariansmatrix som indeholder variansen og kovariansen mellem studierne. Det antages at Σ_i og Ω er uafhængige.

9.4.1 Estimation af ukendte parametre

Hvis det antages at $\widehat{\beta}_i$ og Σ_i er kendte, kan de ukendte parametre β og Ω estimeres ved hjælp af maksimum likelihood estimation eller

restringeret maksimum likelihood estimation. Da det antages at $\widehat{\beta}_i$ er multivariat normalfordelt, er likelihoodfunktionen for $\widehat{\beta}_i$ givet ved:

$$\mathcal{L}(\widehat{\beta}_i) = \prod_{i=1}^N \frac{1}{(2\pi)^{m/2} |\mathbf{\Omega} + \mathbf{\Sigma}_i|^{1/2}} \exp \left(-\frac{1}{2} (\widehat{\beta}_i - \beta)^T (\mathbf{\Omega} + \mathbf{\Sigma}_i)^{-1} (\widehat{\beta}_i - \beta) \right),$$

og log-likelihoodfunktionen er givet ved:

$$\begin{aligned} \ell(\widehat{\beta}_i) = & \sum_{i=1}^N \log \left(\frac{1}{(2\pi)^{m/2}} \right) + \log \left(|\mathbf{\Omega} + \mathbf{\Sigma}_i|^{1/2} \right) \\ & - \frac{1}{2} (\widehat{\beta}_i - \beta)^T (\mathbf{\Omega} + \mathbf{\Sigma}_i)^{-1} (\widehat{\beta}_i - \beta). \end{aligned} \quad (9.4)$$

Log-likelihood funktionen (9.4) kan ikke løses implicit og skal derfor maksimeres ved hjælp af iterative metoder. Det er kun muligt at finde estimatet for β udtrykt ved estimatet $\widehat{\mathbf{\Omega}}$.

$$\overline{\widehat{\beta}}_{\text{tf}} = \left(\sum_{i=1}^N (\widehat{\mathbf{\Omega}} + \mathbf{\Sigma}_i)^{-1} \right)^{-1} \left(\sum_{i=1}^N (\widehat{\mathbf{\Omega}} + \mathbf{\Sigma}_i)^{-1} \widehat{\beta}_i \right)$$

og kovariansmatricen for $\overline{\widehat{\beta}}_{\text{tf}}$ kan findes ved:

$$\text{cov}(\overline{\widehat{\beta}}_{\text{tf}}) = \left(\sum_{i=1}^N (\widehat{\mathbf{\Omega}} + \mathbf{\Sigma}_i)^{-1} \right)^{-1}$$

Til at estimere $\mathbf{\Omega}$ og β anvendes Newton-Raphson for enten $\ell(\widehat{\beta}_i)$ eller $\ell_r(\widehat{\beta}_i)$. Den restringerede log-likelihoodfunktion $\ell_r(\widehat{\beta}_i)$ er givet ved:

$$\ell_r(\widehat{\beta}_i) = \ell(\widehat{\beta}_i) + \log \left(\left| \sum_{i=1}^N \mathbf{\Omega} + \mathbf{\Sigma}_i \right|^{1/2} \right) - m \log \left(\frac{1}{2\pi} \right)$$

Hvilket kan anvendes til at bestemme en konfidensregion for β .

Eksempel 9.4

Betragtes igen problemstillingen fra eksempel 9.3, hvor vi betragter alder og forøget indtag af 10 gram fed fisk i forhold til ændringer i livvidde/højde, anvendes samme estimater samt kovariansmatricer indenfor hvert studie, som ved metaanalyse med fast effekt. Disse ses i Tabel 9.3. Kovariansmatrixen mellem studierne, estimeres ved hjælp af restringerede maksimum likelihood estimation, til at være:

$$\hat{\Omega} = \begin{array}{c|cc} & \text{Alder} & \text{Fed fisk} \\ \hline \text{Alder} & 0,08 & 0,18 \\ \text{Fed fisk} & 0,18 & 0,81 \end{array}$$

Dermed findes det vægtede gennemsnit til:

$$\widehat{\beta}_{\text{tf}} = (-0,26 \ -0,39)^T,$$

hvor første estimat hører til alder og andet estimat hører til 10 gram øget indtag af fed fisk.

Konfidensintervallet for β_{alder} , for fastholdt indtag af fed fisk, bliver da:

$$95\%CI(\widehat{\beta}_{\text{tf}_{\text{alder}}}) = (-0,48; -0,05)$$

og konfidensintervallet for $\beta_{\text{fed fisk}}$, for fastholdt alder, bliver:

$$95\%CI(\widehat{\beta}_{\text{tf}_{\text{fed fisk}}}) = (-1,60; 0,82)$$

Igen ses det at konfidensintervallerne for metaanalyse med tilfældig effekt bliver større end konfidensintervallerne for metaanalyse med fast effekt, ligesom det sås i det univariate tilfælde.]

Ved anvendelse af multivariat metaanalyse kan der opstå problemer med perfekt prædiktion. Perfekt prædiktion medfører at kovariansmatrixen ikke er positive semidefinit, hvilket er nødvendigt for at kunne finde det vægtede gennemsnit med tilhørende konfidensinterval.

9.5 Perfekt prædiktion

Perfekt prædiktion forekommer når alle observationer i en variabel har samme værdi. Eksempelvis, hvis alle observationerne for ét niveau af en kategorisk variabel er én. Tilsvarende kan forekomme ved andre typer af opdeling af data. Ved metaanalyse kan der opstå perfekt prædiktion i nogle studier, men ikke nødvendigvis i andre. Dette skyldes at alle studierne ikke nødvendigvis har information ,for hele udfaldsrummet, for en variabel. Betragtes eksempelvis alder inddelt i aldersgrupperne 20-40, 40-60 og 60+. Hvis et af de studier, der betragtes, kun har deltagere i aldersgrupperne 40-60 og 60+ vil alle værdierne i gruppen 20-40 være nul og dermed opstår der perfekt prædiktion for denne aldersgruppe for dette studie. Hvilket betyder at variationen i den pågældende kategori er nul.[36]

Hvis der detekteres perfekt prædiktion i et studie arrangerer *mmeta_make* data således, at to værdier for hver covariat erstattes med henholdsvis gennemsnittet af covariaten i det specifikke studie plus og minus standardafvigelsen. På denne måde undgås perfekt prædiktion samtidig med, at de tilføjede observationer kun får en lille indvirkning, på resten af modellen.

9.6 Metaanalyse i forhold til blandede modeller

Både metaanalyse og blandede modeller er metoder, der kan anvendes til at modellere ét samlet estimat, for data opdelt i forskellige studier. Begge metoder har deres fordele og ulemper. I det tilfælde hvor man ikke har indviddata, men kun parameterestimerer samt varianskomponenter fra studierne, kan man kun anvende metaanalyse. Metaanalyse blev da også indført til, at analysere denne type af data. Har man adgang til indviddata, har man muligheden for at anvende begge metoder. Ved at anvende blandede modeller opnås en mere kompleks model til modellering af ens problemstilling, fremfor brug af metaanalyse. Man kan derfor spørge sig selv om man så ikke altid bør anvende blandede modeller, såfremt data er til rådighed?

Det der taler imod at anvende blandede modeller er, at det

beregningsmæssigt bliver kompliceret og at variansen for de underlæggende normalfordelinger indgår i estimatberegningen, hvor metaanalysen er baseret på simple beregninger, hvor beregningerne af parameterestimaterne for hvert studie $\hat{\beta}_i$, ikke afhænger af den bagvedliggende variation.

Ved blandede modeller er det ikke nødvendigt at lave et parameterestimat for hver af grupperne, hvorfor der anvendes færre parametre til bestemmelse af den samlede regressions koefficient hørende til eksponeringen. Såfremt man tror på de antagelser man laver ved de blandede modeller, altså at grupperne er normalfordelte og variationen i alle grupperne er ens, får det endelige estimat en mere naturlig fortolkning. Ved anvendelse af metaanalyse med tilfældige effekter, bliver fortolkningen tilnærmelsesvis den samme.

Et punkt hvor de to metoder adskiller sig betydeligt fra hinanden, er ved justering for andre forklarende variabler. Ved metaanalyse, hvor man ikke har data til rådighed, kan det være et problem, at de forskellige studier ikke har justeret for de samme variabler og at effekterne af eksponeringen derfor ikke nødvendigvis er sammenlignelig. Har man alt individdata til rådighed, foretages justeringerne i metaanalyserne for hvert studie separat, hvilket betyder at effekten de forklarende variabler bliver forskellig i de forskellige studier. I de blandede modeller kan man lade de forklarende variabler indgå med fast effekt, hvorved det antages at effekten er ens i alle grupper. Måden man justerer i metaanalysen svarer til, at der indføres interaktion mellem gruppe og alle andre variabler i den blandede model. Tror man på, at der er interaktion, kan man i de blandede modeller teste hvorvidt der er tale om signifikant interaktion og om der eventuelt kun er interaktion mellem nogle af variablerne og gruppe.

Hvilken metode der passer bedst til netop de analyser man skal lave, må derfor blive en vurderings sag. Vi foretrækker blandede modeller da de giver mere fleksible modellerings muligheder, men mener ligeledes at man med fordel kan betragte metaanalyserne, blandt andet til, at vurdere hvorvidt man kan betragte grupperne som en samlet gruppe, eller som flere mindre grupper.

IKKE-LINEÆR MODELLERING AF KONTINUERTE VARIABLER

Efter at have opstillet sin model, skal man tage stilling til, hvordan variablerne i modellen skal indgå. Dette emne belyses med udgangspunkt i [14] og [31].

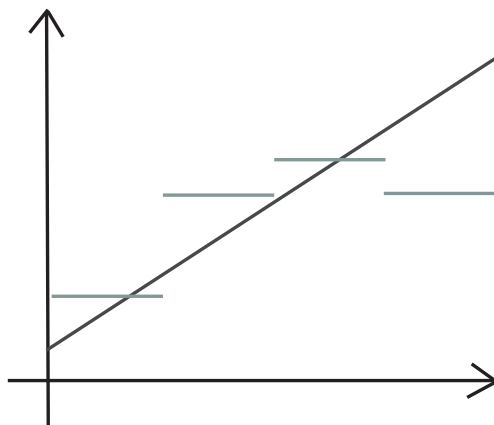
Når kontinuerte variabler medtages i en model, har man som udgangspunkt tre muligheder:

- I** man kan lade variabelen indgå som lineær,
- II** opdelt i kategorier eller
- III** anvende en metode til ikke-lineær modellering af kontinuerte variabler.

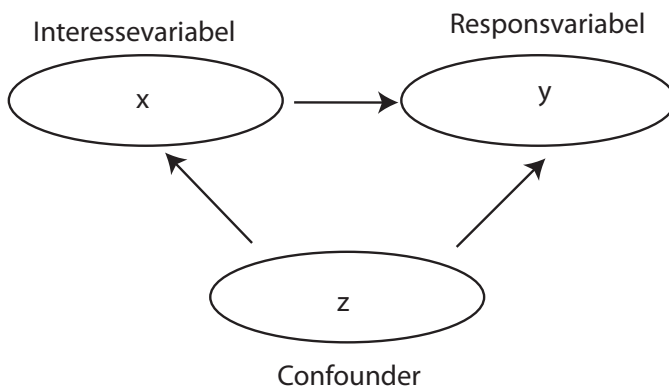
Fordelen ved at lade en variabel indgå som lineær er, at fortolkningen er simpel og der kun skal estimeres en parameter per variabel. For at udnytte mere af datainformation, kan den kontinuerte variabel opdeles i kategorier, og dermed indgå som en ordinal eller nominal variabel. I Figur 10.1 ses forskellen mellem en kontinuert variabel modelleret som lineær og opdelt i fire kategorier.

Ved at opdele en kontinuert variabel i kategorier, opnås en mere detaljeret model, men vigtig information kan gå til spilde [13] og man risikerer derved residualconfounding. Residualconfounding kan medføre, at hele effekten af variabelen ikke bidrager til modellen. Når der justeres for variabelen, fjernes hele effekten altså ikke, svarende til at der indenfor hver kategori, stadig er en afhængighed af variabelen. En confounder er defineret ved at den ikke er et led i årsagskæden fra eksponering til respons, den er en selvstændig årsag til responsen og den er associeret med eksponeringen. En confounder kan illustreres, som det ses i Figur 10.2. Til at illustrere residualconfounding, betragtes eksempelvis den kontinuerte variabel alder og opdelingen af denne i

aldersgrupperne under og over 60 år. Ved at opdele alder i grupper, antager man, at den dag man fylder 60 år, ændrer risikoen sig markant, i stedet for den gradvise ændring, alder ofte har på risikofaktorer.



Figur 10.1: I figuren ses, hvordan mere af data informationen kan udnyttes ved eksempelvis at opdele data i kvartiler, fremfor at modellere data som lineær.



Figur 10.2: I figuren ses, hvordan en confounder kan illustreres. Den ujusterede model vil da være $y \sim x$ og den justerede model, hvor effekten af confounderen ligeledes modelleres vil være $y \sim x + z$.

For at udnytte data bedst muligt, kan ikke-lineær modellering anvendes. Der er udviklet flere forskellige metoder, hvoraf vi vil betragte restringerede kubiske splines og fraktionelle polynomier.

10.1 Restringerede kubiske splines

Restringerede kubiske splines (RCS) er en metode til at inkludere en kontinuert variabel på en glat ikke-lineær måde.

Fremgangsmåden for bestemmelse af restringerede kubiske splines er, at definitionsområdet af den kontinuerte variabel $\mathbf{X}_{(n \times 1)}$ inddeles i $k + 1$ intervaller ved hjælp af k knudepunkter. I hvert interval modellerer et tredjegradspolynomium sammenhængen mellem responsen og den forklarende variabel. Responsen defineres forskelligt, alt efter hvilken type af model der betragtes. Tredjegradspolynomiet restringeres til at være glat i knudepunktet, det vil sige hældningskoefficienten er den samme fra højre og venstre i hvert knudepunkt. Da mængden af data indenfor en variabels definitionsområdet ofte er småt repræsenteret i enderne af definitionsområdet, vil et tredjegradspolynomium i disse områder kunne lave store udsving og derfor bliver spline funktionen uden for de yderste knudepunkter, defineret som lineær. Formen for en restringeret kubisk spline med k knudepunkter $t_1, \dots, t_k$ er givet ved:

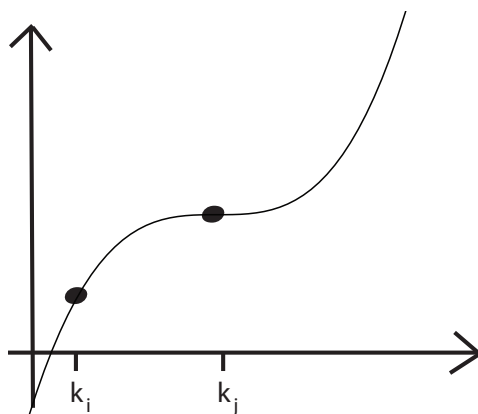
$$f(\mathbf{X}) = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \dots + \beta_{k-1} \mathbf{X}_{k-1}, \quad (10.1)$$

hvor $\mathbf{X}_1 = \mathbf{X}$ og værdien for $\mathbf{X}_j$ for hver deltager i findes ved:

$$x_{ij} = (x_i - t_j)^3 - (x_i - t_{k-1})^3 + \frac{t_k - t_j}{t_k - t_{k-1}} + (x_i - t_k)^3 \frac{t_{k-1} - t_j}{t_k - t_{k-1}}$$

for $i = 1, \dots, n$ og $j = 1, \dots, k - 2$. Notationen $(z)_+$ betyder at ved positive værdier af z anvendes z -værdien og ved negative værdier af z anvendes nul. For hver variabel der modelleres med en begrænset kubisk spline estimeres $k - 1$ parametre, udover interceptet. Det bemærkes, at den lineære model er et specialtilfælde af den restringeret kubisk spline hvor $\beta_2, \dots, \beta_{k-1}$ er nul.

Hvordan tredjegradspolynomiet mellem to knudepunkter k_i og k_j er modelleret er illustreret i Figur 10.3. Mellem de to knudepunkter k_i og k_j anvendes tredjegradspolynomiet, som ses i figuren, mens der udenfor anvendes andre modellerede tredjegradspolynomier eller rette linier, såfremt det er uden for de yderste knudepunkter. Da en restringeret kubisk spline skal være glat i knudepunkterne, skal tredjegradspolynomiet mellem k_{i-1} og k_i skære tredjegradspolynomiet mellem k_i og k_j i punktet k_i samt have samme hældningskoefficient fra venstre side som tredjegradspolynomiet mellem k_i og k_j har fra højre.



Figur 10.3: I figuren ses, hvordan tredjegradspolynomiet mellem to knudepunkter k_i og k_j er modelleret

Desto flere observationer der er i et datasæt, desto større kompleksitet, det vil sige knudepunkter, kan en model bære. Oftest vælges $k = 3, 4, \dots, 7$. Det anbefales af [14] at knudepunkterne som standard, er lokaliseret i faste fraktiler af $\mathbf{X}$, alt efter hvor mange knudepunkter der er. Placeringen af knudepunkterne er vist i Tabel 10.1.

Ved at anvende fraktilernes placering, bestemmes knudepunkterne udelukkende ud fra mængden af data. Der hvor sammenhængen mellem responsvariablen og den forklarende variabel er mest interessant, er ikke nødvendigvis der, hvor der er mest data. Har man i forvejen en ekspertviden, som indikerer en mere interessant eller mere biologisk naturlig placeringen af knudepunkterne, kan denne viden anvendes,

i stedet for fraktilerne, til at fastslå placeringen af knudepunkterne. Betragtes eksempelvis BMI, hvor WHO har opstillet nogle grænser for undervægtig ($\text{BMI} < 18,5$), normalvægtig ($18,5 \leq \text{BMI} < 25$), overvægtig ($25 \leq \text{BMI} < 30$) og kraftig overvægtig ($30 \leq \text{BMI}$). I stedet for at anvende fraktilerne, kan de tre punkter 18,5, 25 og 30 anvendes.

k		Fraktil						
3			0,10	0,50	0,90			
4		0,05	0,35	0,65	0,95			
5		0,05	0,275	0,50	0,725	0,95		
6	0,05	0,23	0,41	0,59	0,77	0,95		
7	0,025	0,1833	0,3417	0,50	0,6583	0,8167	0,975	

Tabel 10.1: I tabellen ses hvorledes knudepunkterne tildeles for $k = 3, 4, \dots, 7$ knudepunkter. [14]

10.1.1 Valg af antallet af knudepunkter

Desto større sammenhæng den forklarende variabel har med responsvariablen, desto større kompleksitet tildeles variabelen. Det vil sige antallet af knudepunkter, en variabel kan tildeles, stiger desto større sammenhæng der er.

Til at vurdere hvor mange knudepunkter splinfunktionen skal anvende, kan flere forskellige metoder anvendes, blandt andet devians differensen og Akaike's informations kriterium, som blev beskrevet i afsnit 8.1.1. Vi har valgt at anvende devians differensen mellem den lineære modellen og modellen indeholdende maksimum likelihood estimatet. Ifølge [14] er Akaike's informations kriterium bedre til, at bestemme antallet af knudepunkter, idet denne også tager højde for modellens kompleksitet. Vi har på trods af dette valgt, at anvende devians differensen for bedre at kunne sammenligne de to metoder restringerede kubiske splines og fraktionelle polynomier, da vi ligeledes anvender devians differensen i forbindelse med fraktionelle polynomier.

Fremgangsmåden for valg af knudepunkter er da som følger:

I Udregn devians differensen $D_{\text{lineær}} - D_{\text{RCS}_k}$ for hver af modellerne fittet med henholdsvis $k = 3, 4, \dots, 7$ knudepunkter.

II Vælg den model som har den højeste værdi af devians differensen.

Hvis der i datasættet findes mange observationer med den samme værdi, således at fraktil-værdierne er sammenfaldene, kan to eller flere af knudepunkterne falde oven i hinanden. I et sådant tilfælde betragtes de sammenfaldene knudepunkter som ét.

10.1.2 Bestemmelse af konfidensbånd

Modelparametrene anvendt til restringerede kubiske splines kan findes ved hjælp af maksimum likelihood estimation. Derfor antages det at splinefunktionen er approksimativt normalfordelt og man kan finde konfidensbåndet på almindelig vis. Konfidensintervallet for et enkelt punkt, med signifikansniveau $\alpha = 0,05$, er givet ved:

$$95\%CI(f(\mathbf{X})) = f(x) \pm 1,96\sigma(f(\mathbf{X})),$$

hvor $\sigma(f(\mathbf{X}))$ er bestemt ved:

$$\sigma(f(\mathbf{X})) = \sqrt{\sigma^2(f(\mathbf{X}))} = \sqrt{\mathbf{X}\Sigma\mathbf{X}^T},$$

hvor Σ er variansmatrixen for β .

Da restringerede kubiske splines ikke er et punkttestimat men en kurve der beskriver forholdet mellem respons- og eksponeringsvariablen, skal konfidensintervallet findes for alle værdier af eksponeringen. På den måde fås et bånd af konfidensintervaller fremfor et enkelt konfidensinterval. Dette bånd kaldes et konfidensbånd eller simultane konfidensintervaller.

Koefficienterne af en restringeret kubiske spline har ingen selvstændig fortolkning. Den samlede funktion for den restringeret kubiske spline skal derfor illustreres grafisk og derefter fortolkes. I STATA 11, som er det statistiske program vi anvender til analyser, er det på nuværende

tidspunkt kompliceret at lave plots af restringerede kubiske splines. Vi har derfor i forbindelse med dette projekt skrevet to tillægs kommander til STATA 11 kaldet *splot_make* og *splot*, som på en simple måde laver spline figurer. Hjælpefilerne for programmerne er at finde i Appendix D og E og programmerne er, at finde på <http://kmd.kvfc.dk/> og installeres i STATA 11 på følgende måde:

Gem filerne i en mappe på din computer og skriv følgende i STATA 11:

```
. net from "sti til mappe"  
  
. net install "kommando navn", replace
```

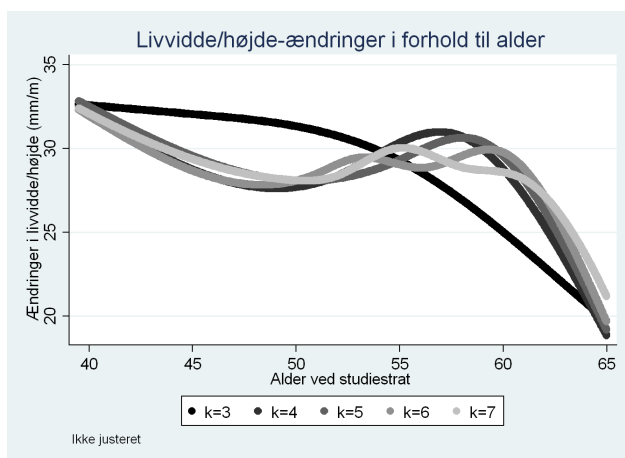
I det følgende eksempel ses det, hvordan restringerede kubiske splines modeller, antallet af knudepunkter bestemmes og konfidensbåndene tilføres.

Eksempel 10.1

Vi betragter responsvariablen ændringer i livvidde/højde og interessevariablen alder. Livvidde/højde-ændringer i forhold til alder udtrykt ved hjælp af en restringeret kubisk spline, med henholdsvis $k = 3, 4, 5, 6$ og 7 , ses i Figur 10.4. Det ses at formen ændrer sig når man anvender $k > 3$ fremfor $k = 3$, men at formen for de resterende splinegrafer minder om hinanden. I Tabel 10.2 er devians differensen udregnet for hver af spline funktionerne, det ses at $k = 6$ og $k = 7$ giver de højeste værdier, hvorfor begge modeller er lige gode. Vi har valgt modellen med $k = 6$ da denne er den mindst komplekse model af de to.

Model	Dev. Diff.	AIC
k=3	210	804
k=4	629	1.221
k=5	653	1.243
k=6	680	1.268
k=7	680	1.266

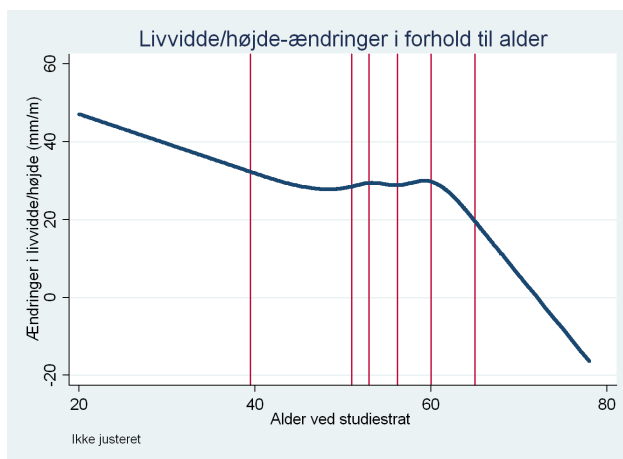
Tabel 10.2: I tabellen ses AIC for modellerne opstillet hvor alder er modelleret som splines med henholdsvis $k=3,4,5,6,7$.



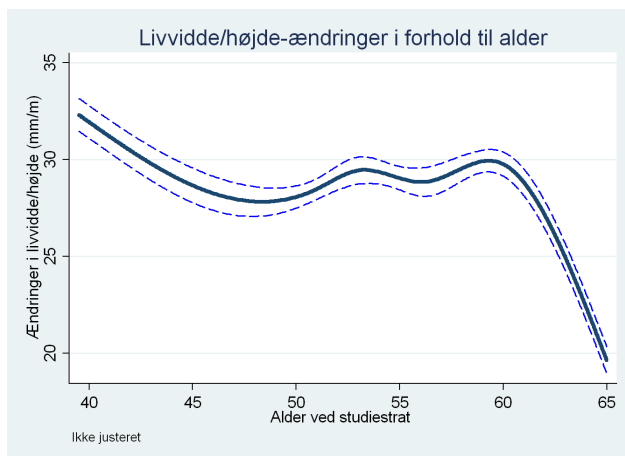
Figur 10.4: I figuren ses spline for alder mod responsvariablen ændringer i livvidde/højde hvor antallet af knudepunkter er henholdsvis $k = 3,4,5,6$ og 7 .

Vi har besluttet at basere vores valg af antallet af knudepunkter på devians differensen. Vi har dog ligeledes i dette eksempel udregnet Akaike's informations kriterium for at se om de to metoder giver den samme model. Akaike's informations kriterium for modellerne med henholdsvis 3,4,5,6 og 7 knudepunkter ses tillige i Tabel 10.2. Ifølge Akaike's informations kriterium, er modellen med $k = 6$ den bedste til at forudsige fremtidige observationer i forhold til modellens kompleksitet, hvilket også var modellen vi valgte udfra devians differensen.

Splinefiguren for livvidde/højde-ændringer i forhold til alder udtrykt ved hjælp af seks knudepunkter, hvor knudepunkternes placering er markeret med lodrette streger, ses i Figur 10.5. Det ses at alle knudepunkterne er lokaliseret mellem 39 og 65 år, idet størstedelen af data er lokaliseret indenfor dette interval. For overskuelighedens skyld vil splinefigurene i resten af rapporten som udgangspunkt kun blive afbilledet indenfor 0,05 og 0,95 fraktilerne. I Figur 10.6 ses splinefiguren, for alder mod responsvariablen ændringer i livvidde/højde med seks knudepunkter, hvor de yderste fem procent ikke er afbilledet og hvor 95 % konfidensbåndet er indtegnet.



Figur 10.5: I figuren ses den restringerede kubiske spline for alder mod responsvariablen ændringer i livvidde/højde med seks knudepunkter.



Figur 10.6: I figuren ses den restringerede kubiske spline for alder mod responsvariablen ændringer i livvidde/højde med seks knudepunkter.

Restringerede kubiske splines ligger meget tæt på data og derfor kan restringerede kubiske splines i visse tilfælde modellere sammenhænge, som biologisk set ikke giver mening. Betragtes eksempelvis resultaterne i Eksempel 10.1 fås at sammenhængen mellem alder og ændringer i livvidde/højde har to små stigninger omkring 54 år og 61 år. At de små stigninger netop sker her har ikke nødvendigvis en biologisk forklaring.

En ulempe ved restringerede kubiske splines er, at valget af antal knudepunkter og placeringen af disse, kan ændre på estimatet af tredjegradspolynomierne [31]. Hvis dette ændrer på effekten af variabelen, kan det ligeledes få konsekvenser for de resultater, der kommer ud af undersøgelsen. Da man ikke har nogen forudsætninger for at fastslå hvilket antal, eller hvilken placering af knudepunkterne, der er mest korrekte, ved man heller ikke hvilken effekt, der er den mest sandfærdige. Ved brug af få knudepunkter, bliver formen af tredjegradspolynomierne mere afhængige af placeringen af knudepunkterne [31], hvorfor det maksimale antal knudepunkter, data tillader i forhold til kompleksitet, bør anvendes.

En anden metode til ikke-lineær modellering af kontinuerte variabler er fraktionelle polynomier. Til forskel fra restringerede kubiske splines har fraktionelle polynomier en forudbestemt form. Resultaterne kan derved ikke variere på samme vis, som ved restringerede kubiske splines.

10.2 Fraktionelle polynomier

Når fraktionelle polynomier (FP) bestemmes, betragtes graden og potensen af det fraktionelle polynomium. Graden bestemmer hvor mange led FP funktionen består af, mens potensen bestemmer FP transformationen. En FP transformation af første grad, en FP1 transformation, er for den i 'te person defineret ved x_i^q , hvor $i = 1, \dots, n$, $x_i > 0$ og potensen $q \in S = \{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3\}$. Hvor transformationen når $q = 0$ defineres til at være den naturlige log-transformation, $\log(x_i)$.

Mængden S er valgt på denne måde for at begrænse antallet af måder FP kan vælges, samt forhindre ekstreme værdier af q . Ekstreme værdier af q kan nemlig medfører, at modellen bliver meget følsom over for afvigende observationer. Potensen q bestemmes således, at den mest optimale transformation indenfor intervallet af S udføres.

Den mest optimale transformation kan bestemmes ud fra flere forskellige kriterier, blandt andet Akaike's informations kriterium beskrevet i afsnit 8.1.1 eller devians differensen. Vi har valgt at anvende devians differensen, idet dette kriterium er implementeret til samme formål i STATA 11's pakke til fraktionelle polynomier.

Hvis den mest optimale transformation i virkeligheden er $q = -1,5$ er denne ikke mulig på grund af det begrænsede S . Indskrænkningen af S er dermed den største begrænsning ved brug af fraktionelle polynomier. Eksperimentelle studier har dog bekræftet, at S er et godt generelt valg af den bedste potens i intervallet $[-2; 3]$ [31, side 79].

Funktionen for det fraktionelle polynomium af første grad er for det i 'te individ, defineret ved:

$$f_{1i}(x_i, q) = \beta_0 + \beta_1 x_i^q.$$

Det bemærkes at den lineære funktion er et specialtilfælde af det fraktionelle polynomium af første grad, hvor potensen $q = 1$.

For hver grad det fraktionelle polynomium øges, indføres endnu en transformation. Et fraktionelt polynomium kan have op til grad m , da vil potensen være givet ved $\mathbf{q} = (q_1 \dots q_m)$. Funktionen for det fraktionelle polynomium vil da være:

$$f_{mi}(x, \mathbf{q}) = \beta_0 + \sum_{j=1}^m \beta_j h_j(x_i),$$

hvor

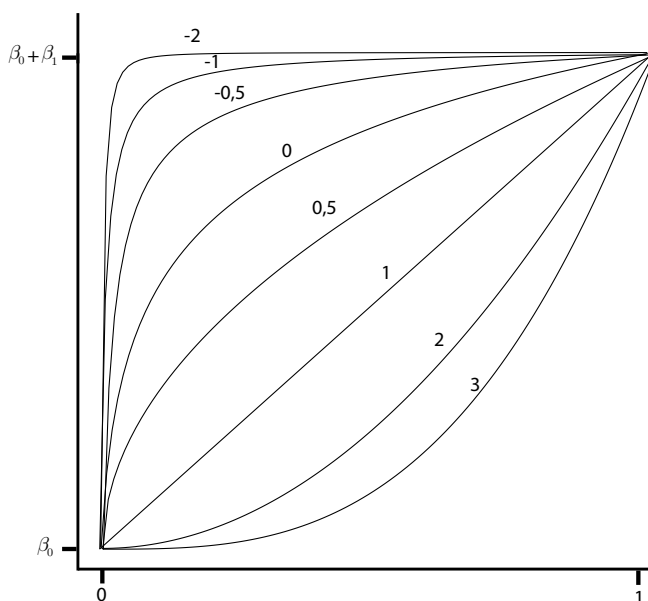
$$h_j(x_i) = \begin{cases} x_i^{q_j} & \text{hvis } q_j \neq q_{j-1} \\ h_{j-1}(x_i) \log(x_i) & \text{hvis } q_j = q_{j-1} \end{cases},$$

hvor $j = 1, \dots, m$ og $i = 1, \dots, n$.

Oftest anvendes det fraktionelle polynomium af første eller anden grad, da højere grader af det fraktionelle polynomium sjældent forbedre model fittet signifikant [31].

Antallet af parametre der skal bestemmes, for et fraktionelt polynomium, er en regressionskoefficient, en værdi af q for hver grad m samt et intercept. For FP2 modellen skal der altså bestemmes $2m+1$ parametre.

Fraktionelle polynomier er en fleksibel klasse af funktioner. Et eksempel på nogle af de former de fraktionelle polynomier kan modellere ses i Figur 10.7 for FP1 og i Figur 10.8 for FP2.



Figur 10.7: I figuren ses eksempler på FP1 kurver.

For Fraktionelle polynomier gælder ligeledes at de:

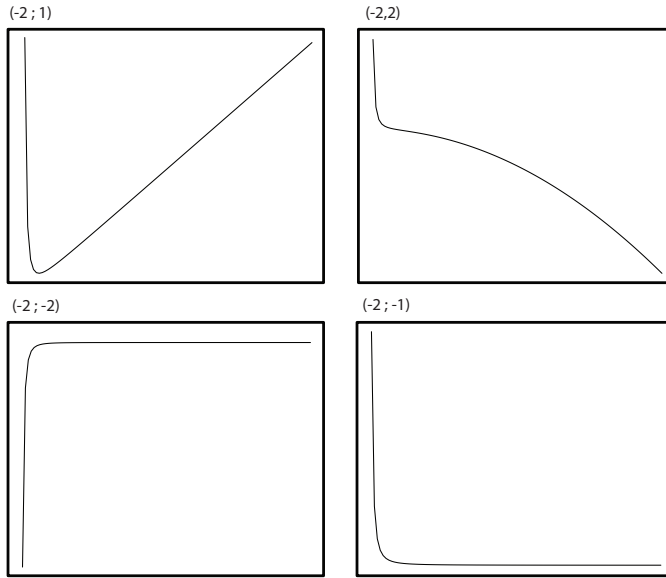
I er asymptotiske når x går mod nul og uendelig.

II har højst et lokalt ekstremum.

III FP1 funktioner er monotone og FP2 er monoton hvis

$\text{sign}(\beta_1 \cdot \beta_2 \cdot q_2) = \text{sign}(q_1)$. Det er ikke muligt at restringere FP2 til at være monoton ved hjælp af maksimum likelihood estimation. Derimod kan man anvende constrained maksimum likelihood estimation som muliggøre restriktion af parameterestimaterne. Da vi i dette projekt ikke har en forhånds antagelse af at sammenhængen mellem ændring i livvidde i forhold til højde og indtag af fed fisk er monoton, vil vi ikke gå dybere ind i constrained maximum likelihood estimation.

Disse egenskaber ved fraktionelle polynomier gør dem egnede til at modellere sammenhænge, hvor man i forvejen har en fornemmelse



Figur 10.8: I figuren ses eksempler på FP2 kurver.

af hvordan sammenhængen vil opføre sig. Det forventes ligeledes, at prædiktion udenfor observationsområdet er god.

10.2.1 Valg af potens samt grad af det fraktionelle polynomium

For at bestemme hvilken potens q der er den optimale, vælges et præcisionsmål. Målet vælges i dette tilfælde, som tidligere nævnt, til at være devians differensen, fundet ud fra maksimum likelihood estimation. Deviance differensen blev defineret i afsnit 8.1.1.

Deviance differensen (8.1), side 75, udregnes mellem den lineære funktion og alle andre mulige valg af S . Det vil for FP1 sige, at den potens $q = i$, hvor $i = -2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3$, som maksimere:

$$D_1 - D_i = -2\ell_{q=1} - 2\ell_{q=i},$$

er den optimale potens.

For FP2 sammenlignes ligeledes med den lineære funktion, det vil sige det fraktionelle polynomium af første grad med $q = 1$.

For FP1 er der otte mulige modeller, og dermed otte valg af potens og for FP2 er der 36 kombinationer af potenser.

Når den optimale potens for hver af graderne er valgt, skal den optimale grad af FP vælges. Til dette opstilles en række tests. Som tidligere nævnt er det oftest tilstrækkeligt at anvende et fraktionelt polynomium af første eller anden grad, hvorfor følgende test kun opstilles for FP1 og FP2. Øges graden, er proceduren tilsvarende. Alle testene udføres på signifikansniveau α ved hjælp af devians test, som er baseret på devians differensen (8.1). Havde q kunnet vælges vilkårligt, ville devians differensen være approximativt χ^2 fordelt med $d.f_i - d.f_j$, hvor $d.f_i$ og $d.f_j$ er antallet af frihedsgrader i de to modeller i og j , som testes mod hinanden. På trods af at q ikke kan vælges vilkårligt antages det, at devians differensen er approximativt χ^2 .

Alt efter om man ønsker en model, der er så kompleks som mulig, eller så simpel som mulig, opstilles testene i forskellig rækkefølge. Hvilken type af model man fortrækker, kommer an på problemstillingen og er en individuel vurdering. Hvor nogle foretrækker en model der er så kompleks som data tillader, foretrækker andre en model, hvor alt der ikke har en signifikant effekt fjernes fra modellen. Når man foretager sin vurdering af hvilken type model man ønsker, er der tre hoved aspekter man skal tage med:

I Modellens evne til at forudsige fremtidige observationer.

II Modellens kompleksitet.

III Forklaringen af modellen.

Første punkt vejer til fordel for, at vælge en så kompleks en model som mulig, da dette som oftest giver den model, der forudsiger fremtidige observationer bedst. Modellens kompleksitet er også et meget vigtigt aspekt, såfremt overfitting er en problemstilling. Modellens

kompleksitet bør således ikke vælges større en troværdighedsmålet tillader, jævnfør afsnit 7.2.1, hvilket taler for en simplere model. Sidste punkt omhandler fortolkning af modellen. I forbindelse med hvilket fraktionelt polynomium der medtages, er der ingen forskel i hvordan modellens fortolkes og dette punkt har derfor ingen betydning for valg af test proces. Modellens fortolkning får dog betydning, når der skal vælges hvilke forklarende variabler der skal medtages i modellen.

For at opnå en så simpel model som mulig, opstilles testene:

I Test den optimale FP2 funktion for $\mathbf{X}$ mod nul-modellen:

$$\beta_0 = \beta_1 = \beta_2 = 0$$

- H_0 : FP2 modellen er ikke bedre end nul-modellen.
- Deviance differensen er approximativt χ^2 med $2m - 0 = 2 \cdot 2 - 0 = 4$ frihedsgrader.

Hvis testet ikke er signifikant stoppes og der konkluderes at effekten af $\mathbf{X}$ ikke er signifikant. Er testet signifikant fortsættes.

II Test den optimale FP2 funktion for $\mathbf{X}$ mod den lineære model.

- H_0 : FP2 modellen er ikke bedre end den lineære model.
- Deviance differensen er approximativt χ^2 med $2m - 1 = 2 \cdot 2 - 1 = 3$ frihedsgrader.

Hvis testet ikke er signifikant stoppes og der konkluderes at effekten af $\mathbf{X}$ kan modelleres som en ret linie. Er testet signifikant fortsættes.

III Test den optimale FP2 funktion for $\mathbf{X}$ mod den optimale FP1 funktion.

- H_0 : FP2 modellen er ikke bedre end FP1 modellen.
- Deviance differensen er approximativt χ^2 med $2m_2 - 2m_1 = 2 \cdot 2 - 2 \cdot 1 = 2$ frihedsgrader, hvor m_2 er graden af FP2 modellen og m_1 er graden af FP1 modellen.

Hvis testet ikke er signifikant stoppes og der konkluderes at effekten af $\mathbf{X}$ kan modelleres som en FP1 funktion. Er testet signifikant anvendes FP2 funktionen.

For at opnå en så kompleks model som mulig, opstilles testene:

I Test den optimale FP2 funktion for $\mathbf{X}$ mod den optimale FP1 funktion.

- H_0 : FP2 modellen er ikke bedre end FP1 modellen.
- Deviance differensen er approximativt χ^2 med $2m_2 - 2m_1 = 2 \cdot 2 - 2 \cdot 1 = 2$ frihedsgrader, hvor m_2 er graden af FP2 modellen og m_1 er graden af FP1 modellen.

Er testet signifikant anvendes FP2 funktionen. Er testet ikke signifikant fortsættes.

II Test den optimale FP1 funktion for $\mathbf{X}$ mod den lineære model.

- H_0 : FP1 modellen er ikke bedre end den lineære model.
- Deviance differensen er approximativt χ^2 med $2m - 1 = 2 \cdot 1 - 1 = 1$ frihedsgrader.

Er testet signifikant anvendes FP1 funktionen. Er testet ikke signifikant fortsættes.

III Test den lineære model mod nul-modellen: $\beta_0 = \beta_1 = 0$.

- H_0 : Den lineære model er ikke bedre end nul-modellen.
- Deviance differensen er approximativt χ^2 med $1 - 0 = 1$ frihedsgrader.

Hvis testet er signifikant stoppes og der konkluderes at effekten af $\mathbf{X}$ kan modelleres som en ret linie. Er testet ikke signifikant konkluderes at der ingen effekt er af $\mathbf{X}$.

Vi vurdere, at såfremt overfitting ikke er et problem i modellen, er modellens evne til at forudsige fremtidige observationer det vigtigste aspekt indenfor modelopstilling, hvorfor vi foretrækker metoden for en så kompleks model som muligt.

10.2.2 Bestemmelse af konfidensbånd

De simultane konfidensintervaller, eller konfidensbånd kan findes som beskrevet i afsnit 10.1.2. Da potensen og graden af det fraktionelle polynomium bestemmes inden parametrene estimeres, skal man være opmærksom på at konfidensintervallerne kan blive for smalle.

I det følgende eksempel ses det, hvordan den optimale potens og grad af et fraktionelle polynomium bestemmes, samt hvordan dette modellers og konfidensbåndet tilføres.

Eksempel 10.2

Vi betragter igen problemstillingen, hvor eksponeringsvariablen er alder og responsvariablen er ændringer i livvidde/højde. Deviance differensen anvendt til at bestemme den optimale potensen for FP1 og FP2, ses i Tabel 10.3. Det ses at devians differensen er størst for FP1 ved $q = 3$ og for FP2 ved $q_1 = 3$ og $q_2 = 3$ og dermed er disse de optimale værdier af potensen for første og anden grad. Efter at have bestemt de optimale potenser, bestemmes graden af det fraktionelle polynomium.

FP1		FP2								
Potens q	Dev.diff	Potens		Dev.diff	Potens		Dev.diff	Potens		Dev.diff
		q_1	q_2		q_1	q_2		q_1	q_2	
-2	-247	-2	-2	-123	-1	1	38	0	2	114
-1	-161	-2	-1	-83	-1	2	92	0	3	165
-0,5	-119	-2	-0,5	-60	-1	3	149	0,5	0,5	62
0	-77	-2	0	-36	-0,5	-0,5	-14	0,5	1	83
0,5	-38	-2	0,5	-9	-0,5	0	6	0,5	2	128
1	0	-2	1	19	-0,5	0,5	27	0,5	3	176
2	73	-2	2	79	-0,5	1	51	1	1	101
3	141	-2	3	142	-0,5	2	102	1	2	143
		-1	-1	-51	-0,5	3	156	1	3	189
		-1	-0,5	-32	0	0	24	2	2	180
		-1	0	-10	0	0,5	44	2	3	219
		-1	0,5	13	0	1	66	3	3	253

Tabel 10.3: I tabellen ses devians differensen for alle mulige valg af q for FP1 og FP2. Devians differenserne, som er markerede er de højeste indenfor hver grad.

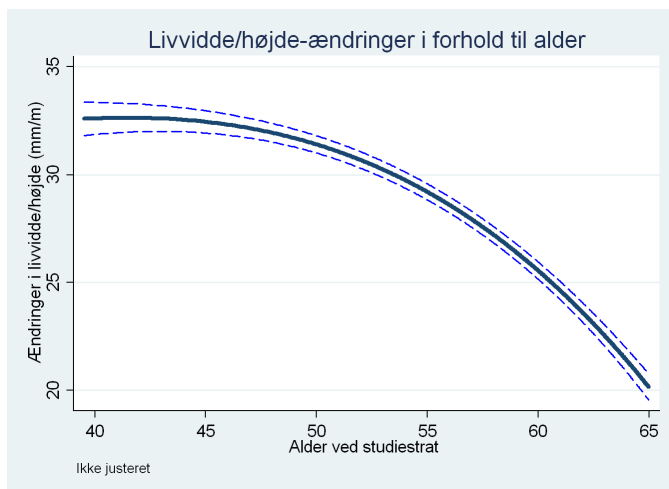
For at bestemme hvilken grad, der vælges foretages ovenstående procedure for så kompleks en model som muligt. Testresultaterne ses i Tabel 10.4. Det ses at graden af det mest optimale fraktionelle polynomium er $m = 2$. Testet for så simpel en model som muligt, giver samme resultat. Det fraktionelle polynomium for alder mod livvidde/højde-ændringer, hvor $m = 2$, $q_1 = 3$ og $q_2 = 3$, har for det i 'te individ forskriften:

$$\begin{aligned}f_2(x_i, \mathbf{q}) &= \beta_0 + \beta_1 x_i^3 + \beta_2 x_i^3 \log(x_i) \\&= 0,0295 + 0,0004 x_i^3 - 0,0002 x_i^3 \log(x_i)\end{aligned}$$

og ses med 95 procent konfidensbånd i Figur 10.9. Parameterestimaterne er fundet ved hjælp af maximumlikelihood estimation, som beskrevet i Kapitel 8.

Model	Deviance	d.f	$\mathbf{q}$	Sammenligning	Dev. diff.	P
FP2	-252535	4	3,3	FP2 mod FP1	112	< 0,001
FP1	-252423	2	3	-	-	-

Tabel 10.4: I tabellen ses resultaterne af devians testene for bestemmelse af grad af det fraktionelle polynomium.



Figur 10.9: I figuren ses et fraktionelt polynomium for alder mod responsvariablen livvidde i forhold til højde, hvor $m = 2$, $q_1 = 3$ og $q_2 = 3$.

10.3 Diskussion

Vi har nu gennemgået flere forskellige metoder til modellering af kontinuerte variable, hvoraf vi har haft hovedvægt på ikke-lineær modellering. Begge metoder gennemgået i forbindelse med ikke-lineær modellering har deres fordele og ulemper.

De restringerede kubiske splines anvender $k - 1$ frihedsgrader, mens de fraktionelle polynomier anvender $2m$ frihedsgrader. I de fleste tilfælde, vil de restringerede kubiske splines bruge flere frihedsgrader end de fraktionelle polynomier. Er modellen for kompleks, i forhold til data, kan det derfor være en fordel, at anvende fraktionelle polynomier eller få knudepunkter, for at undgå overfitting.

Da begge metoder anvender mere end en parameter til, at modellere en variabel, skal multivariat metaanalyse anvendes, hvis der ønskes at lave metaanalyser af resultaterne. Da restringerede kubiske splines er mere fleksible end fraktionelle polynomier, forventes det, at resultaterne

af en multilevel metaanalyse for restringerede kubiske splines bliver svære at fortolke, mens resultaterne for fraktionelle polynomier vil give en fortolkelig graf. For begge metoder er det essentielt at formen af forskriften er ens for alle de studier der betragtes, altså at antallet af knudepunkter er ens for de restringerede kubiske splines og at potensen samt graden er ens for de fraktionelle polynomier.

Restrangerede kubiske splines ligger sig tæt op af data, hvilket kan resultere i at biologisk uforklarlige fænomener, kan forekomme. Dette kan være en ulempe, hvis de restringerede kubiske splines skal anvendes til at forklare sammenhængen mellem ens responsvariabel og interessevariabel. Skal de restringerede kubiske splines derimod anvendes til, at fjerne effekt fra confoundere, er det en fordel, at de modellere så tæt ved data. På den måde fjernes mere effekt af den pågældende variabel i netop det datasæt man betragter.

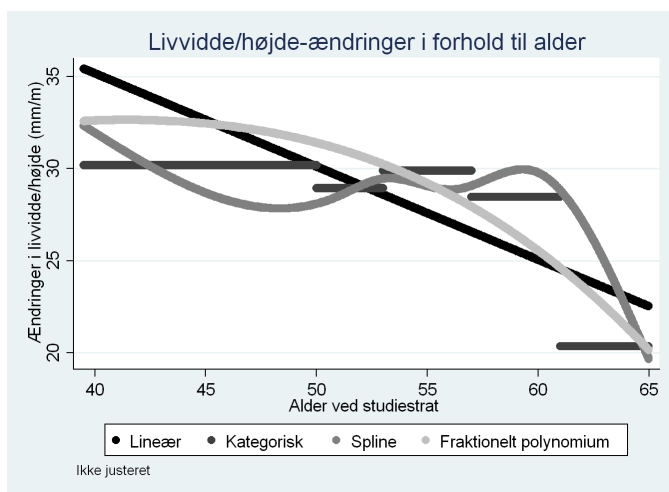
De fraktionelle polynomier giver pæne asymptotiske kurver og kan restringeres til at være monotone. På den måde undgår man biologiske uforklarlige hændelser. Har man en forhåndsviden om, at den variabel man betragter er konstant aftagende, er det en fordel at kunne indbygge det i modellen. Er man derimod på helt bar bund med hvordan ens variabel opfører sig, kan restringerede kubiske splines mere frit give et billede af variabelernes sammenspil.

Ønsker man at udtale sig om sammenhængen mellem to variabler, udenfor observationsområdet, giver restringerede kubiske splines en utilstrækkelig prædiktion udenfor observationsområdet, da modellen her er fittet med rette linier som kun er baseret på området mellem yderste kundepunkt og maksimum eller minimums observationen. Mens fraktionelle polynomier forventes at give en god prædiktion udenfor observationsområdet.

Ved restringerede kubiske splines kan antallet af knudepunkter og deres placering ændre på resultatet af ens model, hvilket kunne tale for at anvende fraktionelle polynomier. Ved fraktionelle polynomier vælger man, ikke desto mindre, sin model udfra et test, hvor antagelserne ikke er opfyldt. Ifølge [31] kan dette gøres op for ved, at anvende et højere

signifikansniveau α .

Lige gyldig hvilken metode man vælger at anvende er det, hvis sammenhængen mellem en responsvariabel Y og en forklarende variabel X er ikke-lineær, bedre end at anvende en lineær model eller at konvertere data til en kategoriske variabel. I Figur 10.10 betragtes resultaterne for det gennemgående eksempel vedrørende alders sammenhæng med ændringer i livvidde/højde som henholdsvis lineær, opdelt i kvintiler, modelleret som splines med seks knudepunkter og modelleret som et fraktionelt polynomium med $m = 2$ og $p_1 = p_2 = 3$. Alle modellerne viser en faldende effekt med alderen, men er ikke enige om udviklingen henover alder.



Figur 10.10: I figuren ses modellerne som modellere alders sammenhæng med ændringer i livvidde/højde som henholdsvis lineær, opdelt i kvintiler, modelleret som splines med seks knudepunkter og modelleret som et fraktionelt polynomium med $m = 2$ og $p_1 = p_2 = 3$.

En måde at skelne mellem hvornår man med fordel kan anvende den ene metode fremfor den anden, er om man har forhåndsviden eller ej. Har man forhåndsviden kan fraktionelle polynomier anvendes, mens restringerede kubiske splines anvendes, hvis der ingen forhåndsviden er.

En anden måde at skelne mellem hvornår man med fordel kan anvende den ene metode fremfor den anden, er hvilken type variabel man betragter. Hvis variabelen er en forklarende variabel som medtages for at fjerne confounder effekt, fjerner restringerede kubiske splines mest af effekten. Er variabelen interessevariabel og man eventuelt ønsker at se hvad der sker udenfor observationsområdet, kan fraktionelle polynomier med fordel anvendes.

I kapitel 9 gennemgik vi multivariat metaanalyse som er en metode til, at kombinere flere assosiationsmål fra flere studier til et samlet assosiationsmål. Multivariat metaanalyse kan derfor ligeledes anvendes til at bestemme et samlet estimat af koefficienterne for henholdsvis restringerede kubiske splines og fraktionelle polynomier. Koefficienterne i sig selv har ingen fortolkning. Kun den samlede funktion kan illustreres grafisk og derefter fortolkes. På nuværende tidspunkt findes der, ifølge vores viden, ingen statistiske programmer, som kan lave en sådan grafisk illustration. Vi har derfor konstrueret henholdsvis et program til grafisk illustration af fraktionelle polynomier, samt et tilsvarende program for restringerede kubiske splines i STATA 11. Hjælpefilerne til de to programmer findes i Apendix F og G og programmerne kan findes på <http://kmd.kvfc.dk/>. For at installere programmerne med hjælpefiler hentes filerne fra hjemmesiden ind på tilsvarende vis som for beskrevet i afsnit 10.1.

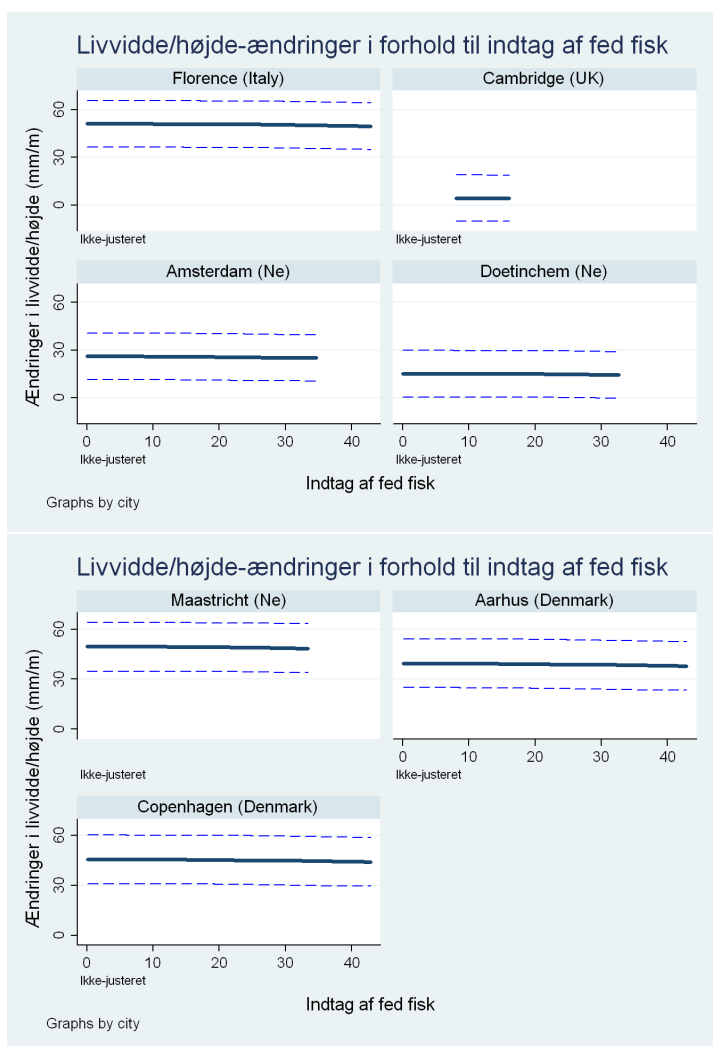
Appendix

NOTATION

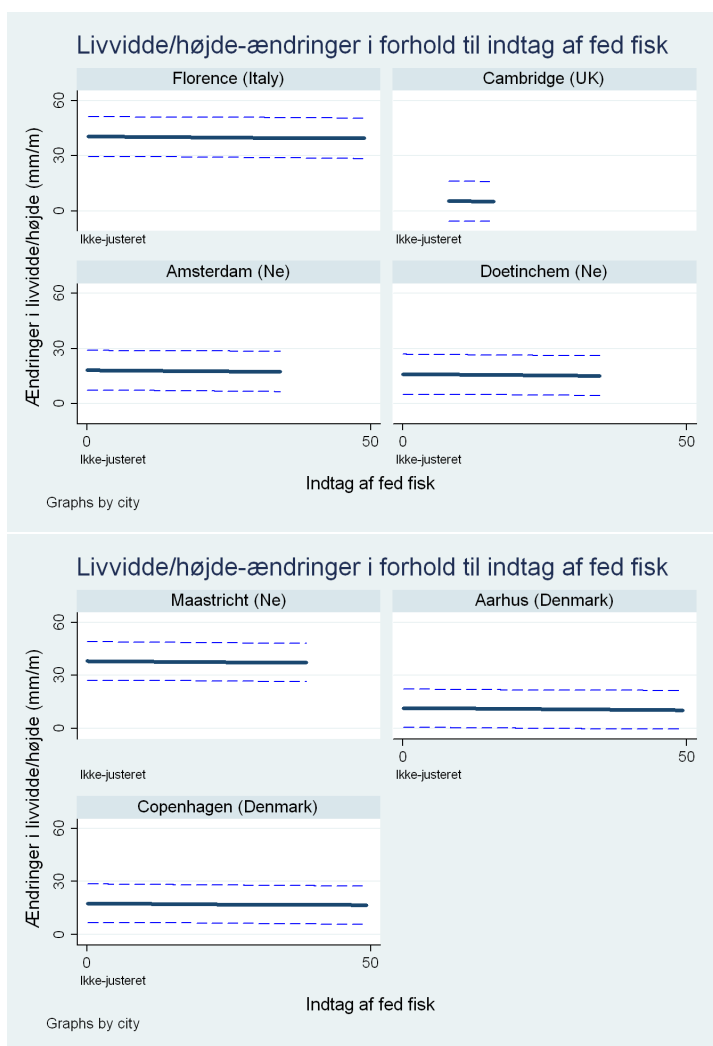
$\mathbf{Y}$	Vektor af responsvariablen.
$\mathbf{X}$	Matrix af forklarende variabler hørende til de faste effekter.
$\mathbf{Z}$	Matrix af forklarende variabler hørende til de tilfældige effekter.
β	Vektor bestående af regressionskoefficienter for de faste effekter.
$\mathbf{b}$	Vektor bestående af modelparametrene for de tilfældige effekter.
$\mathcal{L}(\cdot)$	Likelihoodfunktion.
$\ell(\cdot)$	Log-likelihoodfunktion.
$\mathcal{L}_r(\cdot)$	Restringeret likelihoodfunktion.
$\ell_r(\cdot)$	Restringeret log-likelihoodfunktion.
$S(\cdot)$	Scorefunktion.
$I(\cdot)$	Fischers information.
D	Devians.
LR	Likelihoodratio
$\hat{\cdot}$	Estimat.
$\overline{\beta}_{fa}$	Vægtet gennemsnit for univariat metaanalyse med faste effekter.
$\overline{\beta}_{tf}$	Vægtet gennemsnit for univariat metaanalyse med tilfældige effekter.
$\overline{\beta}_{fa}$	Vektor af vægtede gennemsnit for multivariat metaanalyse med faste effekter.
$\overline{\beta}_{tf}$	Vektor af vægtede gennemsnit for multivariat metaanalyse med tilfældige effekter.
$f(y; \cdot)$	Tæthedsfunktion for y .
$\mathcal{N}$	Normalfordelingen.
χ^2	Chi i anden fordelingen.
$\sim$	Fordelt som.
$E(\cdot)$	Middelværdi.
σ^2	Varians.

w_i	Vægt for det i 'te studie.
$\mathbf{W}_i$	Vægt matrix for det i 'te studie.
Σ_i	Kovarians matrix for det i 'te studie.
Ω	Kovarians matrix mellem studier.
$\mathbf{X}^T$	$\mathbf{X}$ transponeret.
$\mathbf{X}^{-1}$	Invers af $\mathbf{X}$.
$\text{sign}(\cdot)$	Fortegn.
$\mathbb{I}$	Indikatorvariabel.
Δ	Ændring.
n	Antal observationer.
N	Antal grupper.
p	Antal modelparametre.
t_k	Det k 'te knudepunkt for en restringeret kubisk spline.
k	Antal knudepunkter.
q	Potens af et fraktionelt polynomium.
m	Graden af det fraktionelle polynomium
α	Signifikansniveau.

CENTERSPECIFIKKE RESULTATER - KVINDER



CENTERSPECIFIKKE RESULTATER - MÆND



HJÆLPEFIL FOR SPLOT_MAKE

Hjælpefil for programmet *splot_make*. Installationsfiler for programmet findes på <http://kmd.kvfc.dk/>. Programmet er skrevet til STATA 11, men virker også i ældre versioner.

help splot_make,

Title

splot_make — Prepares the data before creating restricted cubic spline plots.

Syntax

Syntax for **splot_make**

```
splot_make varname , [adjcont(varnames)  
    ref(varname=reference value) sref(real) nknots(integer)  
    knots(numlist) n(integer)]
```

Description

splot_make Prepares the data before creating spline plots.

The function appoints the reference values for which the spline plot is to be drawn.

Options

Optionals

adjcont(varnames) specifies the continuous variables one wish to adjust for in the model.

The command **adjcont**() sets the reference value as the median value.

If one wish to define the reference value oneself the command

ref() has to be used.

If the continuous variable is represented in the model it is necessary to represent it in either **adjcont()** or **ref()**

adjdisc(varnames) specifies the discrete variables one wish to adjust for in the model.

The command **adjdisc()** sets the reference group as the lowest level of the variable.

If one wish to define the reference value oneself the command **ref()** has to be used.

If the discrete variable is represented in the model it is necessary to represent it in either **adjdisc()** or **ref()**

ref(varname=value) sets the reference value by your choice.

The command **ref()** makes it possible to set the reference value for the variables for which one do not wish to use the default reference value. The command applies for both continuous and discrete variables.

sref(real) sets the reference value for the spline variable one wish to draw.

If the value entered in **sref()** is less than the value of the second knot it specifies the reference value. If one wish to specify another reference value it is at this moment not possible in this program.

If **sref()** is not specified the reference value is by default zero.

nknots(integer) specifies the number of knots.

The default location of the knots is given by:
Harrell, F.E: Regression Modeling Strategies with Applications to
Linear Models, Logistic Regression and Survival Analysis.
New York: Springer-Verlag 2001. Page 23.

knots(numlist) specifies the location of the knots.

The number of knots must at least be three and the values of these knots must be given in increasing order.

n(integer) specifies how detailed the spline figure is.

The default value of **n**() is 20. If the degree of detail in the figure is too low increase the value.

NB

. The function creates the spline variables `_S'varname'*` which is to be used in the following model.

. To make the final plot the function has to be followed by an regression and the command *splot*.

Examples

```
. webuse lbw
```

```
splot_make bwt
```

```
xi: regress age _Sbwt*
```

```
splot bwt
```

```
. webuse lbw
```

```
splot_make bwt, adjdisc(race) ref(lwt=129) nknots(7)
```

```
xi: regress age _Sbwt* i.race lwt
```

```
plot bwt, legend(off) xtitle(Birthweight) ///  
ytitle(Age of mother)
```

Also see

help *splot*

HJÆLPEFIL FOR SPLOT

Hjælpefil for programmet *splot*. Installationsfiler for programmet findes på <http://kmd.kvfc.dk/>. Programmet er skrevet til STATA 11, men virker også i ældre versioner.

help splot,

Title

splot — Draws restricted cubic spline plots.

Syntax

Syntax for splot

```
splot varname , [legend([contents] [location]) exp
ytitle(name of y-axis) xtitle(name of x-axis) title(title of plot)
note(note) name(name[, replace])]
```

Description

splot Draws spline plots with confidence bands.

Options

Optionals

legend([contents] [location]) Option for specifying legend.

To see legend options also see **legend_option**

exp takes exp to the predicted values and their confidence intervals.

ytitle(name of y-axis) specify the title to appear next to the y-axis.

xtitle(name of x-axis) specify the title to appear next to the x-axis.

title(title of plot) specifies the overall title of the plot.

The title appears centred at the top of the plot.

note(note) specifies notes to be displayed with the plot.

Notes are displayed in a small font placed at the bottom-left corner of the plot.

name(name[, replace]) specifies the name of the plot.

If **name**() is not specified, **name**(Graph, replace) is assumed.

NB

opt **splot** requires the command **help splot** and an regression to be. run additionally

. **splot** requires the command *splot_make* and an regression to be run additionally.

Examples

```
. webuse lbw
```

```
splot_make bwt
```

```
xi: regress age _Sbwt*
```

```
splot bwt
```

```
. webuse lbw
```

```
splot_make bwt, adjdisc(race) ref(lwt=129) nknots(7)
```

```
xi: regress age _Sbwt* i.race lwt
```

```
splot bwt, legend(off) xtitle(Bitrhweight) ///
```

ytitle(Age of mother)

Also see

help *splot_make*

HJÆLPEFIL FOR MVFRAC

Hjælpefil for programmet *mvspine*. Installationsfiler for programmet findes på <http://kmd.kvfc.dk/>. Programmet er skrevet til STATA 11 og virker ikke i ældre versioner.

help mvfrac,

Title

mvfrac — Draws fractional polynomial plots after multivariate metaanalysis.

Syntax

Syntax for mvfrac

mvfrac varname, **power**(power of the fractional polynomial)
by(the variable who defines the different studies)
dataset(the original dataset) [**conf constant**
ytitle(name of y-axis) **xtitle**(name of x-axis) **title**(title of plot)
note(note) **name**(name[, replace])]

Description

mvfrac Draws fractional polynomial plots for each study as well as the weighted average found by multivariate metaanalysis. If the option **conf** is chosen the weighted average with 95 % confidence intervals is drawn without the fractional polynomial plots for each study.

Options

Prescribed

power(# #) specify the power of the fractional polynomial.

Both fractional polynomials of first and second degree can be

specified.

If the fractional polynomial is of first degree the option **power** is specified as **power**(#)

If the fractional polynomial is of second degree the option **power** is specified as **power**(# #)

by(variable) specify the variable who defines the different studies.

dataset(dataset) specify the original dataset containing the variable of interest.

Optionals

coef

If **coef** is specified the output of the program *mvfrac* is the fractional polynomial found by using multivariate metaanalysis with 95 % confidence intervals.

constant specifies that the constant is also of interest.

ytitle(name of y-axis) specify the title to appear next to the y-axis.

xtitle(name of x-axis) specify the title to appear next to the x-axis.

title(title of plot) specifies the overall title of the plot.

The title appears centred at the top of the plot.

note(note) specifies notes to be displayed with the plot.

Notes are displayed in a small font placed at the bottom-left corner of the plot.

name(name[, replace]) specifies the name of the plot.

If **name()** is not specified, **name(Graph, replace)** is assumed.

NB

The function has to be run after *mvmeta_make* and *mvmeta*.

The fractionel polynomium has to be made by *fracgen* or be called 'name'_1 for degree one and 'name'_1 and 'name'_2 for second degree. If the name of the variable is more than six digits the function collapses. Because of the way *fracgen* defines the variable names.

Examples

```
. webuse lbw, clear
```

```
fracgen bwt .5 3
```

```
xi: mvmeta_make regress age bwt_*, ///  
saving(mvlbw) replace by(race) names(b v) esave(N)
```

```
use mvlbw, clear
```

```
mvmeta b v
```

```
mvfrac bwt, power(.5 3) by(race) dataset(lbw.dta) ///  
constant
```

Also see

help *mv spline*, *mvmeta_make* and *mvmeta*

HJÆLPEFIL FOR MVSPLINE

help mvspline,

Title

mvspline — Draws restricted cubic spline plots after multivariate metaanalysis.

Syntax

Syntax for mvspline

mvfrac varname, **by**(the variable who defines the different studies)
dataset(the original dataset) [**knots**(numlist)
nknots(number of knots) **conf** **constant**
ytitle(name of y-axis) **xtitle**(name of x-axis)
title(title of plot) **note**(note) **name**(name[, replace])]

Description

mvfrac Draws restricted cubic spline plots for each study as well as the weighted average found by multivariate metaanalysis. If the option **conf** is chosen the weighted average with 95 % confidence intervals is drawn without the fractional polynomial plots for each study.

Options

Prescribed

by(variable) specify the variable who defines the different studies.

dataset(dataset) specify the original dataset containing the variable of interest.

Optionals

knots(# # ... #) specifies the exact location of the knots.

The values of the knots must be given in increasing order.

nknots(#) specify the number of knots.

The number of knots specified by **nknots** must be between 3 and 7.

The default number of knots when neither **nknots** nor **knots** is given is 5.

coef

If **coef** is specified the output of the program *mvfrac* is the fractional polynomium found by using multivariate metaanalysis with 95 % confidence intervals.

constant specifies that the constant is also of interest.

ytitle(name of y-axis) specify the title to appear next to the y-axis.

xtitle(name of x-axis) specify the title to appear next to the x-axis.

title(title of plot) specifies the overall title of the plot.

The title appears centred at the top of the plot.

note(note) specifies notes to be displayed with the plot.

Notes are displayed in a small font placed at the bottom-left corner of the plot.

name(name[, replace]) specifies the name of the plot.

If **name()** is not specified, **name(Graph, replace)** is assumed.

NB

The function has to be run after *mvmeta_make* and *mvmeta*.

The splines has to be made by *rc_spline* ore be called
_S'name'1, _S'name'2, _S'name'3 and os on.

Examples

```
. webuse lbw  
rc_spline bwt
```

```
xi: mvmeta_make regress age _Sbwt*, ///  
saving(mvlbw) replace by(city) names(b v)///  
esave(N)
```

```
use mvlbw, clear
```

```
mvmeta b v
```

```
mvspline bwt, by(race) dataset(lbw.dta) constant
```

Also see

help *mvfrac*, *mvmeta_make* and *mvmeta*

LITTERATUR

- [1] J. M. Berg, J. L. Thymoczko, and L. Stryer. *Biochemistry*. Freeman, 5. edition, 2002.
- [2] S. Borzoei, M. Neovius, B. Barkeling, A. Teixeira-Pinto, and S. Rossner. A comparison of effects of fish and beef protein on satiety in normal weight men. *European Journal of Clinical Nutrition*, 60(7):897–902, 2006.
- [3] J. M. Bruun and T. Christiansen. Overvægt og forhøjet kolesteroltal (dyslipidæmi), 2009. <http://www.netdoktor.dk/overvaegt/overvaegtogkolesterol.htm>.
- [4] E. Calle, C. Rodriguez, K. Walker-Thurmond, and M. J. Thun. Overweight, Obesity, and Mortality from Cancer in a Prospectively Studied Cohort of U.S. Adults. *The New England Journal of Medicine*, 348(17):1625–1638, April 2003.
- [5] R. C. Cheng. Bootstrapping Simultaneous Confidence Bands. *Winter Simulation Conference*, (37):240–247, Juni 2005.
- [6] B. D. Cox, M. J. Whichelow, and A. T. Prevost. The development of cardiovascular disease in relation to anthropometric indices and hypertension in British adults. *International Journal of Obesity*, 22(10):966–973, Oktober 1998.
- [7] R. Dersimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188, 1986.
- [8] DioGenes. Methodological Guideline DIOGENES WP3.1. (2), Marts 2007.
- [9] K. M. Due. Livsstilsfaktorerens betydning for livvidde. *Institut for Matematiske Fag - Aalborg Universitet*, 9. semester, December 2009.
- [10] EPIC. Panacea website (27/11-09), 1993. <http://epicnl.eu/English/Introduction/tabid/444/Default.aspx>.

- [11] Fødevarestyrelsen. De 8 kostråd, 2010. http://www.altomkost.dk/NR/rdonlyres/1B1875C2-FF4C-42FF-B874-DEE858B4EE77/0/02_Kostraad_2_fisk09.ppt,.
- [12] Danmarks Fiskeriforening, Danske Fiskeres Producent Organisation, Danish Seafood Association, Danmarks Fiskehandlere, Konsumfiskeindustriens Arbejdsgiver Forening, and Dansk Akvakultur. Ernæring og sundhed - Næringsværdi skema, 2010. http://www.2gangeomugen.dk/fileadmin/downloads/diverse/N_ringsv_rdier.pdf,.
- [13] S. Greenland. Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis. *Epidemiology*, 6(4):356–365, Juli 2005.
- [14] F. E. Harrell Jr. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*, Edition = , Year = 2001, ISBN = 978-0-387-95232-1. Springer.
- [15] D. A. Harville. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *American Statistical Association*, 72(358):329–338, Juni 1977.
- [16] D. W. Haslam and W. P. James. Obesity. *The Lancet*, 366:1197–1209, Juli 2005.
- [17] S. Y. Ho, T. H. Lam, and E. D. Janus. Waist to stature ratio is more strongly associated with cardiovascular risk factors than other simple anthropometric indices. *Annals of Epidemiology*, 13(10):683–691, November 2003.
- [18] S. H. Holt, J. C. Miller, P. Petocz, and E. Farmakalidis. A satiety index of common foods. *European Journal of Clinical Nutrition*, 49:675–690, 1995.
- [19] S. D. Hsieh and T. Muto. The superiority of waist-to-height ratio as an anthropometric index to evaluate clustering of coronary risk factors among non-obese men and women. *European Journal of Clinical Nutrition*, 40(2):216–220, Februar 2005.

- [20] R. Huxley, S. Mendis, E. Zheleznyakov, and S. Reddy J. Chan. Body mass index, waist circumference and waist:hip ratio as predictors of cardiovascular risk-a review of the literature. *European Journal of Clinical Nutrition*, 64(1):16–22, August 2009.
- [21] G. Leandro. *Meta-analysis in Medical Research*. BMJ, 2005.
- [22] L. Madsen, B. Liaset, and K. Kristiansen. Macronutrients and obesity: views, news and reviews. *Future Medicine*, 3(1):43–74, Februar 2008.
- [23] Danmarks Lungeforening. Rygning (20/12-2007), 2007. <http://www.lungeforening.dk/rygning>.
- [24] J. A. Milewicz, U. Tworowska, and M. Demissie. Menopausal obesity, myth or fact? *Climacteric*, 4(4):273–839, 2001.
- [25] P. Olofsson. *Probability, Statistics, and Stochastic Processes*. Wiley, 2005.
- [26] J. C. Pineiro and D. M. Bates. *Mixed-effects models in S and S-plus*. Springer, 2000.
- [27] J. Reilly. Physical activity and obesity in childhood and adolescence. *The Lancet*, 366(9482):268–269, Juli 2005.
- [28] J. J. Reilly. Physical activity and obesity in childhood and adolescence. *The Lancet*, 366(9482):268–269, 2005.
- [29] E. Riboli, K. J. Hunt, N. Slimani, P. Ferrari, T. Norat, M. Fahey, U. R. Charrondière, B. Hémon, C. Casagrande, J. Vignat, K. Overvad, A. Tjønneland, F. Clavel-Chapelon, A. Thiébaud, J. Wahrendorf, H. Boeing, D. Trichopoulos, A. Trichopoulou, P. Vineis, D. Palli, H. B. Bueno de Mesquita, P. H. M. Peeters, E. Lund, D. Engeset, C. A. González, A. Barricarte, G. Berglund, G. Hallmans, N. E. Day, T. J. Key, R. Kaaks, and R. Saracci. European Prospective Investigation into Cancer and Nutrition (epic): study populations and data collection. *Public Health Nutrition*, 5(6B):1113–1124, 2002.

- [30] R. D. Riley, K. R. Abrams, P. C. Lambert, A. J. Sutton, and J. R. Thompson. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*, 26(1):78–97, Marts 2007.
- [31] P. Royston and W. Sauerbrei. *Multivariable Model-Building - A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley, 2008.
- [32] L. Sjöström, A. Lindross, M. Peltonen, J. Torgerson, C. Bouchard, B. Carlsson, S. Dahlgren, B. Larsson, K. Narbro, C.D. Sjöström, M. Sullivan, and H. Wedel. Lifestyle, Diabetes, and Cardiovascular Risk Faktors 10 Years after Bariatric Surgery. *The New England Journal of Medicine*, 351(26):2683–2693, December 2004.
- [33] A. J. Stunkard and T. I. Sørensen. Obesity and Socioeconomic Status - A Complex Relation. *The New England Journal of Medicine*, 329(14):1036–1037, September 1993.
- [34] A. M Uhe, G. R. Collier, and K. O’Dea. A comparison of the effects of beef, chicken and fish protein on satiety and amino acid profiles in lean male subjects. *The Journal of Nutrition*, 122:467–472, 1992.
- [35] H. C. van Houwelingen, L.R. Arends, and T. Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21(4):589–624, 2002.
- [36] I. R. White. Multivariate random-effects meta-analysis. *The Stata Journal*, 9(1):40–56, 2009.
- [37] WHO. Controlling the global obesity epidemic (27/11-09), 2009. <http://www.who.int/nutrition/topics/obesity/en/>.
- [38] A. Yafune, T. Funatogawa, and M. Ishiguro. Extended information (EIC) approach for linear mixed effects model under restricted maximum likelihood (REML) estimation. *Statistics in Medicine*, 24(22):3417–3429, 2005.