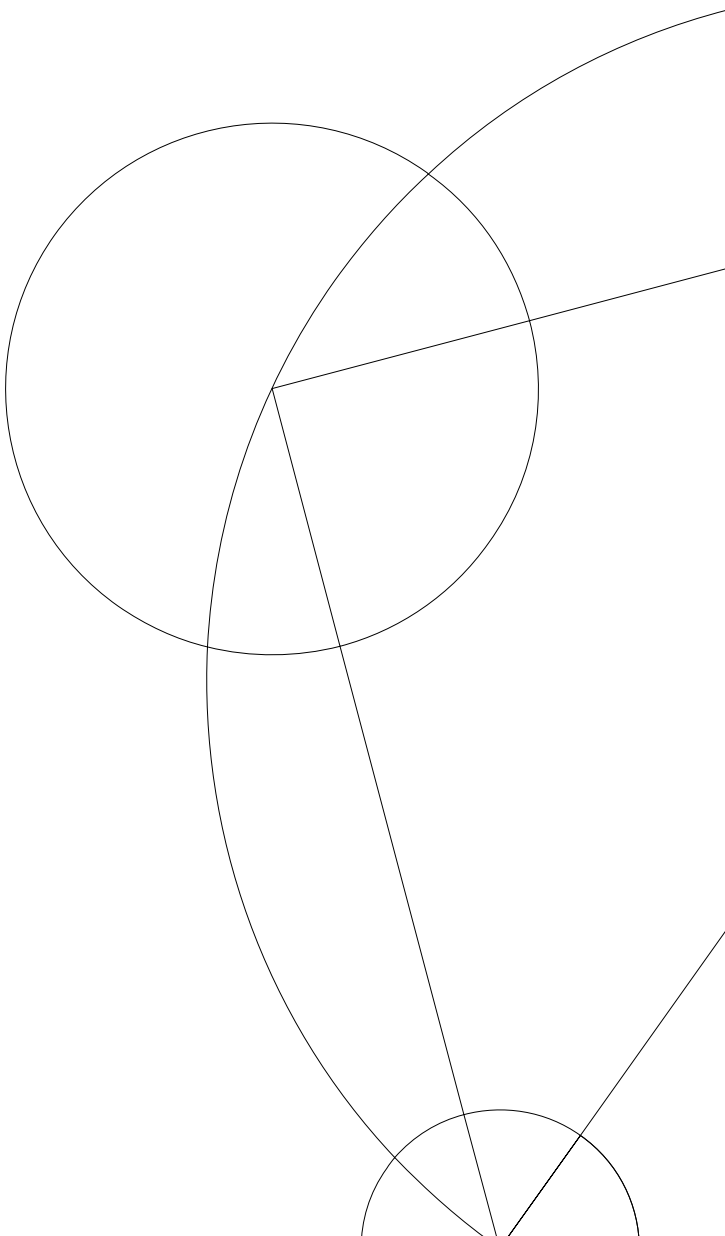
KØBENHAVNS UNIVERSITET

Eksamensprojekt

Sabine V.L. Hansen, Olivia T.G. Helmersen, Mathias F. Jensen og Simon K. Harmat

Løn- og kønsfordeling på universitetsuddannelser

Social Data Science



26. august 2016

Eksamensopgavens dele er besvaret af følgende gruppemedlemmer:

- **Sabine V.L Hansen:**
  - Indledning
  - Modeller: Validering
- **Olivia T.G. Helmersen:**
  - Data, metode og etik
  - Modeller: Regressionstræ
- **Mathias F. Jensen:**
  - Analyse
  - Modeller: Logit
- **Simon K. Harmat:**
  - Konklusion
  - Modeller: Machine learning

Indholdsfortegnelse

Indledning

Investeringer i human kapital sker med forventninger til afkast i arbejdsmarkedet (Becker 1964). Grundet direkte samt indirekte diskrimination, i.e. lønforskelle, kønnene imellem, er kønnenes traditionelle specialisering, mænd i lønnet arbejde og kvinder i hjemmet, fremhævet som økonomisk optimal (Becker 1991; Becker 1985). Med forventning om denne kønsspecialisering forventes det yderligere, at kvinder investerer mindre i human kapital, i.e. løndiskrimination er cirkulær og selvforstærkende fra et human kapitalsynspunkt (Blackburn et al. 2002). På trods af den forudsete cirkularitet, er der gennem de seneste årtier sket dramatiske kønsforandringer i human kapital investeringer: Kvinder udgør nu majoriteten af studerende på lange videregående uddannelser (se side 5). Igen, fra et human kapital perspektiv burde kvinders øgede investeringer i human kapital betyde mindre lønforskelle mellem kvinder og mænd. Markante lønforskelle kønnene imellem er dog fortsat observeret, endda også imellem kvinder og mænd med lange videregående uddannelser, hvor mænd i 2014 havde en bruttoindkomst 36,54 % højere end kvinder, jf. Tabel 1.

Table 2: Bruttoløn fordelt på køn, 2014

	Mænd	Kvinder
Lange videregående uddannelser	670.133	490.782

Kilde: Danmarks Statistik

En mulig forklaring på disse lønforskelle er divergerende afkast på investeringer i forskellige lange uddannelser. Uddannelsesvalg er præget af segregering: Kun få videregående uddannelser har et optag af studerende med lige andele af mænd og kvinder. Kønsforskelle i uddannelsesvalg kunne skyldes divergerende præferencer kønnene imellem (Hakim 2000), men *'... individual preferences (and thus choices) are always socially embedded and constrained, and may be shaped by unjust background conditions, as well as by habit and engrained normative assumptions.'* (Crompton 2007: 234). Socialt determinerede kønspræferencer i forhold til uddannelse har potentiale til at cementere lønforskelle kønnene imellem på trods af samme resulterende niveau af human kapital. Derfor bør det undersøges, hvorvidt der foreligger en tendens blandt kvinder til at søge mod lange videregående uddannelser, der leder til relativt lavere lønninger, og en tendens blandt mænd til at søge mod lange videregående uddannelser, der resulterer i relativt højere lønninger.

Hvis uddannelser med overrepræsentation af kvinder generelt leder til lavere lønninger, kan det både skyldes kønsforskelle i socialt afgrænsede præferencer, i.e. kvinder søger aktivt mod disse lavere lønnet uddannelser, eller lønforskelle opstået grundet feminisering af bestemte faggrupper. Øget feminisering af faggrupper resulterer ofte i lavere lønninger for disse grupper, grundet for eksempel diskrimination eller over-crowding (Rubery 2015; Bergmann 1974). Sagt på en anden måde, kan lavere lønninger for kandidater fra uddannelser med en overrepræsentation af kvinder skyldes, at kvinder søger mod uddannelser med lavere lønninger, samt at kvinder generelt modtager lavere løn, og lønniveauet for uddannelser med overrepræsentation af kvinder derfor er faldet. Uanset forklaringen, er det klart at lønulighed kønnene imellem allerede initieres ved kvinder og mænds valg af uddannelse.

Som analysen i opgaven viser, kan uddannelsesvalg alene forklare omkring en fjerdedel af lønforskellen blandt kvinder og mænd med lange videregående uddannelser. Dette kan skyldes en række faktorer, blandt andet kønspræferencer for uddannelsesretninger. Specifikke uddannelsesretninger leder til arbejde i højtlønnede faggrupper, såsom mandsdominerede business og tekniske uddannelser. Uddannelsesretninger kan forklare en del af lønforskellen kvinder og mænd imellem, men selv inden for disse specifikke uddannelsesretninger segregeres kønnene – mænd søger også her i højere grad mod uddannelser med højere lønafkast.

Med kønsopdelt data for ansøgerantal samt optagne studerende på alle lange videregående uddannelser i Danmark er det muligt at bestemme omfanget af kønssegregering i uddannelsesvalg. Opgavens centrale problemstilling er at bestemme, hvilken rolle forventede lønforskelle imellem kvindelige og mandelige kandidater spiller for kønsfordelingen på lange videregående uddannelser. Denne analyse foretages ved hjælp af løndata for færdige kandidater

Data, metode og etik

På Uddannelses- og Forskningsministeriets hjemmeside findes en række offentlig tilgængelige datasæt der beskriver forskellige parametre for universitetsoptag. Til opgavens analyse bruges henholdsvis datasæt for ansøgere og optagne fordelt på køn, samt adgangskvotienter. På hjemmesiden kan disse oplysninger hentes for forskellige år. For at automatisere datalæsningen konstrueres en såkaldt

scraper, som tager følgende to argumenter; i) hvilken type datasæt og ii) ønskede antal år fra den nyeste observation. Til opgaven hentes data for årene 2013-2016 for at sikre en stor, repræsentativ stikprøve, samt for at påvise eventuelle nylige forandringer i uddannelsesvalg. Dermed henter funktionen 12 forskellige datasæt.

Løndata opgjort per uddannelse for alle færdige kandidater er produceret af CEPOS og ligger ligeledes offentligt tilgængeligt. CEPOS har udarbejdet et datasæt for top 250 gennemsnitlige bruttolønninger opgjort efter færdiggjort lang videregående uddannelse for alle danskere mellem 25-59 år, men disse data er ikke opdelt efter køn. Da disse løndata er gennemsnitlige og inkluderer både mandlige og kvindelige kandidater fra de respektive uddannelser, mistes dimensionen af kønsforskelle i løn inden for de individuelle uddannelser. På den anden side gør dette løngennemsnit det muligt at isolere den del af lønforskelle, der opstår alene som resultat af kønsforskelle i uddannelsesvalg, hvilket netop er formålet med denne analyse.

De to datakilder er forskellige i den forstand, at universiteternes antal ansøgninger, optag og adgangskvotienter er opgjort for bacheloruddannelser, men løndata for færdige kandidater, i.e. for færdige lange videregående uddannelser. I denne analyse indgår derfor kun bacheloruddannelser, der direkte leder til en kandidatuddannelse, e.g. gennem retskrav på optagelse på en specifik kandidatuddannelse. De to datasæt er derfor forenet ved at identificere par af bachelor- og kandidatuddannelser, e.g. bachelor- og kandidatstudierne i medicin. Dette udelukker for eksempel professionsbacheloruddannelser, der kunne lede til en kandidatuddannelse. Ydermere, introduceres en bias, da sammenhængen mellem parrene af bachelor- og kandidatuddannelser ikke er fuldstændig, e.g. grundet frafald og uddannelsesskift. En anden databegrænsning er, at der i løndata også findes en lille række brede kategorier, som samler flere uddannelser, e.g. uspecficeret civilingeniører. Ydermere identificeres kvalitativt 10 uddannelsesretninger, såsom sundhedsvidenskab og samfundsvidenskab. Denne kategorivariabel er essentiel for en vurdering af uddannelsesretningers indflydelse på lønforskelle.

Den manuelle parring af løn og ansøgere, samt kategoriseringen af uddannelsesretninger, foretages for et enkelt år af data. Til alle uddannelser hører et unikt optagelsesnummer, som gør det muligt at forene alle 12 datasæt, således at løn og uddannelsesretninger programmeres for de restende år.

De anvendte data er offentligt tilgængelige, derfor forefindes ikke rettighedsproblemer. Etisk kan det problematiseres, at nogle uddannelser har meget få optagne, især efter opdeling af køn, således at det er muligt at identificere enkelte studerende i det endelige datasæt. Da ingen data er på individniveau, hverken i datasættene eller i analysen, er etiske problemer derfor ikke eksisterende.

Analyse

For perioden 2013-16 viser data fra de danske universiteter klart, at kvinder nu udgør majoriteten af både ansøgere (54,22 %) og optagne (53,10 %) på bacheloruddannelser, der giver direkte adgang til en kandidatuddannelse. På trods af dette, findes der på arbejdsmarkedet fortsat store lønforskelle kønnene imellem, og uddannelsesvalg alene kan forklare en stor del af disse lønforskelle. Ved hjælp af antallet af optagne kvinder og mænd på de individuelle uddannelser samt forventede lønninger efter endt uddannelse, er det muligt at forudsige gennemsnitlige lønninger for optagne kvinder og mænd i perioden 2013-16.

Table 3: Vægtede gennemsnitlige forventede lønninger fordelt på køn, 2013-16

	2013	2014	2015	2016
Kvinder	567.313	568.562	572.960	572.933
Mænd	612.272	614.439	617.768	619.243
Forskel	44.959	45.877	44.809	46.310

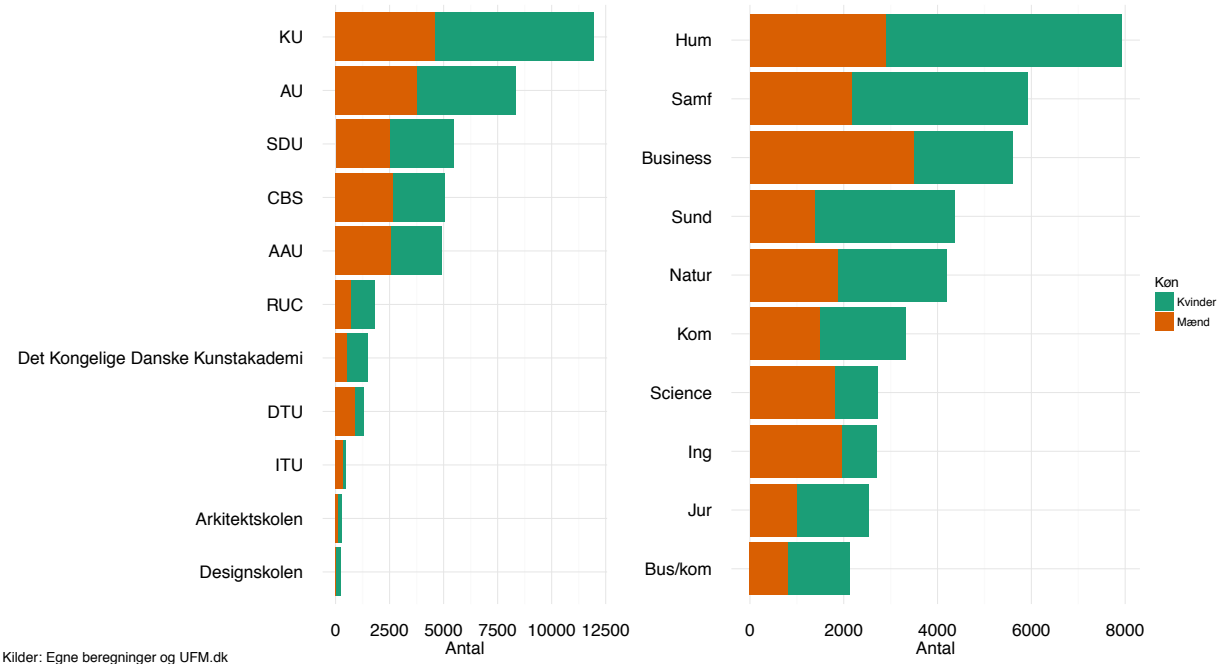
Kilder: Egne beregninger, UFM og CEPOS

De vægtede gennemsnit er vist i Tabel 2. Der er stor forskel i de forudsagte gennemsnitslønninger for kvinder og mænd, og uddannelsesvalg alene har derfor stor betydning for senere lønforskelle på arbejdsmarkedet. Ydermere viser Tabel 2, at de forudsagte lønforskelle har været meget stabile gennem de seneste fire år.

Nedenfor viser Figur 1 de absolutte antal af kvindelige og mandlige ansøgere fordelt på de relevante

uddannelsesinstitutioner. Kvinder udgør majoriteten på de fleste traditionelle universiteter, men billedet er anderledes på CBS, DTU og ITU. Desuden kan det også ses, at uddannelsesretninger som ingeniør, science og business er domineret af mandlige ansøgere. Dette giver sig også til udtryk i de gennemsnitlige antal af kvindelige og mandlige ansøger på det højest lønnede uddannelser, Figur 2.

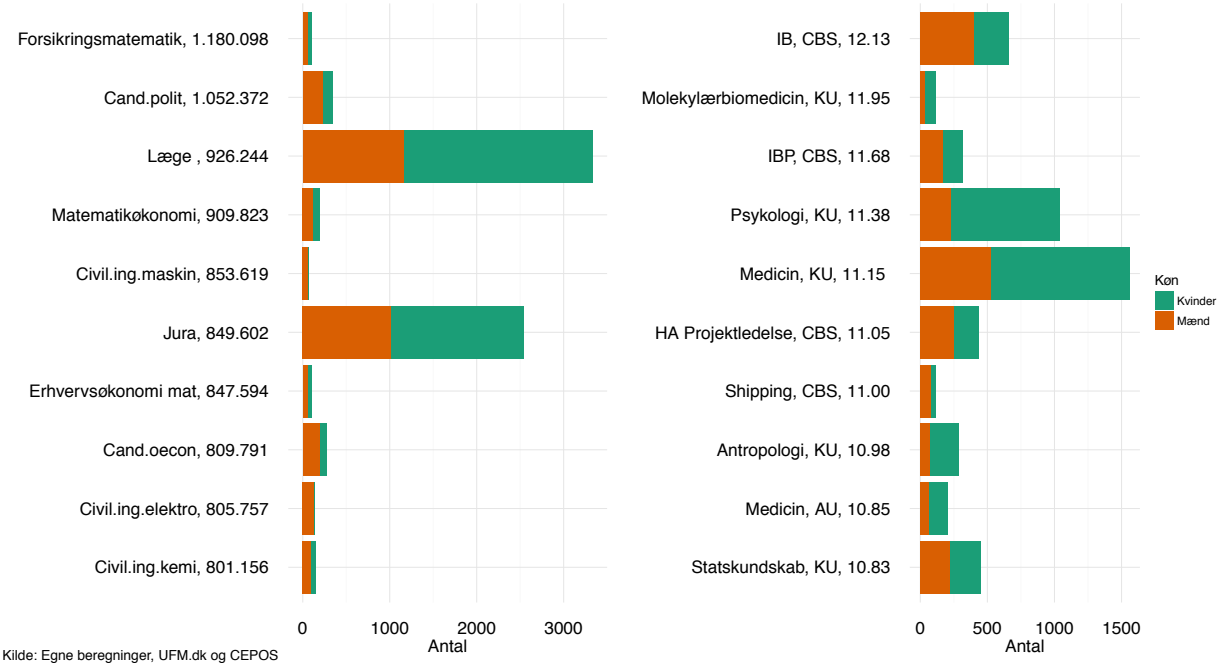
Figur 1: Gennemsnitlig antal ansøgere fordelt på køn ved danske uddannelse institutioner og retninger, 2013-16



Kilder: Egne beregninger og UFM.dk

De højest lønnede uddannelser er alle dominerede af mandlige ansøgere, med undtagelse af to: professionerne lægevidenskab og jura. Dette replicerer det velkendte – og internationale – billede af kvinders koncentration på lange videregående uddannelser, der giver adgang til professioner (Connolly and Gregory 2007). Inden for professionerne, såsom lægevidenskab, jura og psykologi, er det klart defineret, hvilke niveauer af human kapital, der er nødvendige for at praktisere. I professionerne er kvinder og mænd med de samme kvalifikationer derfor i høj grad ansat i lignende stillinger – der er mindre plads til diskrimination (Connolly and Gregory 2007).

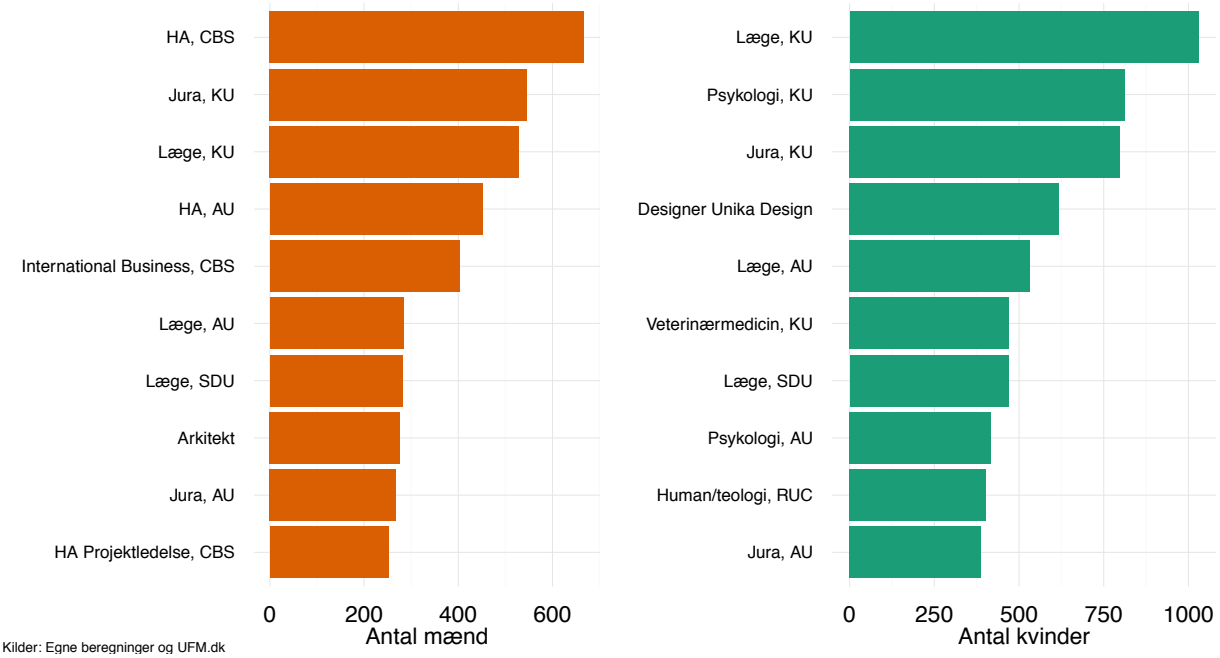
Figur 2: Gennemsnitlige antal ansøgere fordelt på køn for top 10 uddannelser målt på hhv. bruttoløn (kr.) og højeste adgangskvotienter, 2013-16



Kilde: Egne beregninger, UFM.dk og CEPOS

Denne søgen mod professionerne afspejler sig tydeligt i kvinders uddannelsesvalg, både i forhold til de top 10 mest søgte uddannelser og i forhold til de 10 uddannelser med de højeste adgangskvotienter, Figur 3. Her ses en tydelig tendens: Mænd søger mod højtlønnede businessuddannelser, hvor kvinder i langt højere grad søger mod professionerne lægevidenskab, psykologi, jura samt veterinærmedicin.

Figur 3: De 10 mest søgte uddannelser for kvinder og mænd, 2013-16



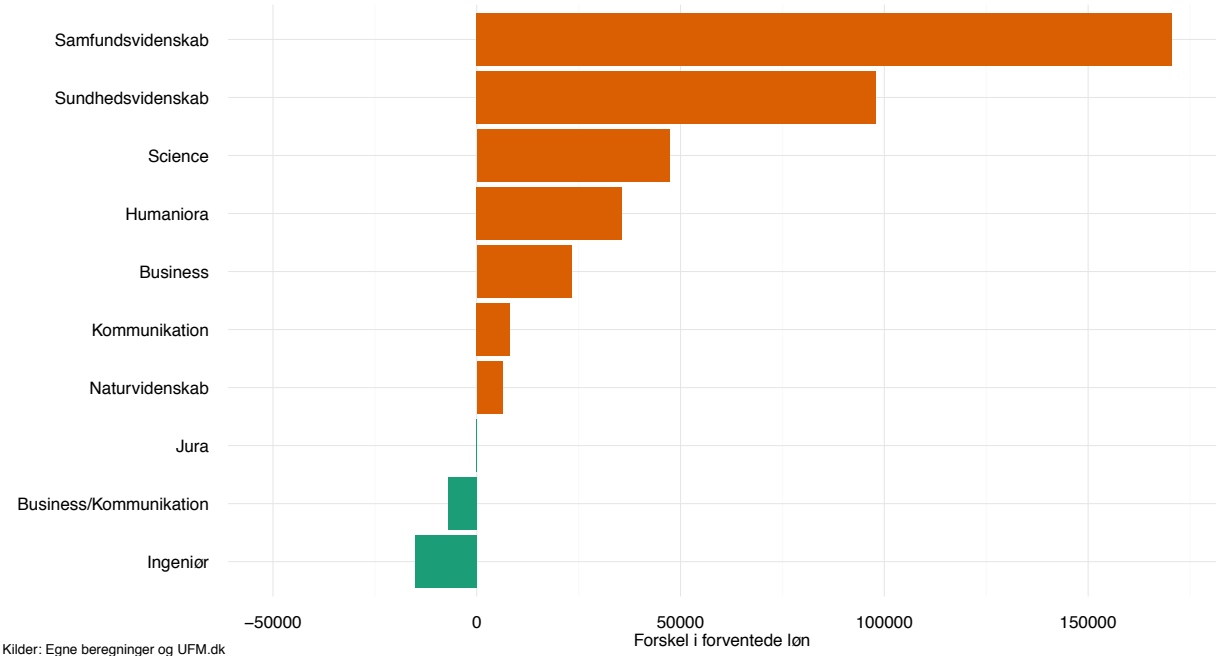
Kilder: Egne beregninger og UFM.dk

Figur 2 viser ligeledes, at de 10 uddannelser med højest lønnede kandidater alle er tekniske og matematiske, med undtagelse af de to professioner, lægevidenskab og jura, hvor kvinder udgør majoriteten af ansøgerne. Mænds dominans i de højest lønnede tekniske og matematiske fag, beviser at kønsstereotype antagelser bliver cementeret i både i uddannelsesvalg og på arbejdsmarkedet: Traditionelle, mandsdominerede faggrupper oplever højere lønninger, og disse faggrupper tiltrækker en stor overvægt af mandelige ansøgere – lønulighed er dermed med en cirkulær proces. Grundet store lønforskelle på tværs uddannelsesretninger, e.g. ingeniørvidenskab og jura, kan lønforskelle kønnene imellem derfor potentielt forklares gennem divergerende kønspræferencer for uddannelsesretninger, e.g. flere mænd søger mod de højtlønnede uddannelsesretninger business og ingeniørvidenskab. På den anden side er det muligt, at mænd også koncentrerer sig i de højest lønnede uddannelser inden for de forskellige uddannelsesretninger, og uddannelsesretninger dermed kun er en mindre del af forklaringen.

Ifølge Figur 4 er dette faktisk tilfældet: Med undtagelse af business/kommunikation samt ingeniørvidenskab koncentrerer mænd sig i de bedst lønnede uddannelser, også inden for specifikke uddannelsesretninger. Dermed er kønsspecialisering og -koncentration i specifikke uddannelsesret-

ninger kun en delvis forklaring af lønforskelle mellem kvinder og mænd med lange videregående uddannelser.

Figur 4: Forskel i forventede, vægtede lønninger fordelt på uddannelsesretning



For at finde et mere robust billede af kønsforskelle i uddannelsesvalg, introduceres udvidede modeller i det følgende afsnit.

Modeller

I et forsøg på at forklare kønsfordelingen på lange videregående uddannelser udregnes andelen af kvinder optaget på alle de forskellige lange videregående uddannelser i Danmark. Disse kvindeandele benyttes herefter som afhængig variabel i en række økonometriske modeller. Først benyttes en logit-regressionsmodel, da den afhængige variabel, kvindeandele er proportionel. Dernæst anvendes en statistisk læringsmodel til at opstille et regressionstræ for processen bag de kønsspecifikke uddannelsesvalg. Dermed vurderes både kausale effekter, samt forudsigelseskraften af de uafhængige variable.

Logit

Da variablen, der beskriver kvindeandelen, er proportional, falder dens værdier mellem 0 og 1. Dermed bør forudsagte værdier også falde i dette interval. Dette opnås ved at anvende en generaliseret lineær model (glm) med et logit link og en binomial familie, som er at foretrække, selv med en kontinuert afhængig variabel (Baum 2008). Estimerede logit-koefficienter er ikke lig med effektstørrelse, som ved almindelige lineære estimeringer. Derfor udregnes mere informative gennemsnitlige marginale effekter (Average Partial/Marginal Effects, APE) efter estimeringer af logit-modellerne, og disse rapporteres i Tabel 3. Ved estimering vægtes observationerne efter uddannelsesstørrelse, i.e. antal optagne på studierne. I modellen vægtes uddannelser dermed proportionelt med deres størrelse. Formålet med logit-modellen er at bestemme kausale sammenhænge mellem kvindeandelen på uddannelser og en række andre variable.

Der opstilles fire modeller, hvori der indgår forskellige uafhængige variable. I den første model estimeres udelukkede effekten på kvindeandelen af forventet gennemsnitlig indkomst efter endt uddannelse. Det skal her bemærkes, at indkomst kan være endogen, altså afhængig af kvindeandelen, hvormed der kan være omvendt kausalitet. I model 2 introduceres en kategorisk variabel, der beskriver de 10 forskellige uddannelsesretninger, hvilket tydeligvis er relevant, jf. Figur 4. Kategorien business/kommunikation anvendes som reference kategori. De næste to modeller inkluderer to yderligere kontinuerte variable: Model 3 medregner adgangskvotienten på de forskellige uddannelser, og Model 4 inddrager dernæst uddannelsesstørrelsen, i.e. det totale optag på studierne. Det fulde model med alle fire variable skrives således:

$$logit(kvindeandel_i) = \alpha_i + \beta_1 \cdot indkomst_i + \beta_2 \cdot business_i + \beta_3 \cdot humaniora_i + \beta_4 \cdot ingenioer_i + \beta_5 \cdot jura_i$$

$$+\beta_6 \cdot kommunikation_i + \beta_7 \cdot naturvidenskab_i + \beta_8 \cdot samfundsvidenskab_i + \beta_9 \cdot science_i$$

$$+\beta_{10} \cdot sundshedsvidenskab_i + \beta_{11} \cdot adgangskvotient_i + \beta_{12} \cdot totaloptag_i + \epsilon_i$$

De estimerede logit-koefficienter fremgår i appendix, Tabel A1, medens de estimerede APE kan ses her, i Tabel 3:

Tabel 3: Logit-modeller, Average Partial Effects

Model	(1)	(2)	(3)	(4)
Indkomst i 100.000 kr.	-0.035***	-0.042***	-0.040***	-0.039***
	(0.001)	(0.001)	(0.001)	(0.001)
Business		-0.174***	-0.172***	-0.157***
		(0.008)	(0.008)	(0.008)
Humaniora		-0.042***	-0.018*	-0.016*
		(0.008)	(0.008)	(0.008)
Ingeniør		-0.291***	-0.264***	-0.270***
		(0.009)	(0.009)	(0.009)
Jura		0.109***	0.077***	0.111***
		(0.010)	(0.010)	(0.011)
Kommunikation		-0.085***	-0.065***	-0.067***
		(0.008)	(0.009)	(0.009)
Naturvidenskab		-0.057***	-0.034***	-0.036***
		(0.008)	(0.008)	(0.008)
Samfundsvidenskab		-0.016*	-0.019*	-0.011
		(0.008)	(0.008)	(0.008)
Science		-0.295***	-0.264***	-0.264***
		(0.009)	(0.009)	(0.009)
Sundhedsvidenskab		0.184***	0.142***	0.158***
		(0.010)	(0.010)	(0.010)
Adgangskvotient			0.011***	0.011***
			(0.000)	(0.000)
Totalt optag i hundrede				-0.009***
				(0.001)

Kilder: Egne beregninger, UFM.dk og CEPOS. Standardfejl i parenteser.

Model 1 viser en klar sammenhæng mellem kvindeandel og indkomst; en sammenhæng som er både statistik og økonomisk signifikant. Der er en klar tendens til at andelen af kvinder på uddannelser falder, når forventede lønninger stiger. Model 1 viser, at for hver 100.000 kroner forventede lønninger stiger, falder andelen af kvinder med 3,5 %. Model 2 inddrager uddannelsesretninger, som tydeligvis også er afgørende for kvindeandelen på uddannelser. Business, ingeniørvidenskab og science er domineret af mænd, medens jura og sundhedsvidenskab er domineret af kvinder. Overraskende forstærkes effekten af indkomst på kvindeandele efter introduktionen af uddannelsesretninger i modellen, og nu falder andelen af kvinder med 4,2 % per 100.000 kroner forøgelse i forventet gennemsnitlig indkomst. Model 3 viser, at uddannelsernes adgangskvotient har en lille positiv sammenhæng med kvindeandelen på studierne – kvinder lader til at have højere karakterer i deres adgangsgivende eksamen, eller i hvert fald udnytte disse karakterer ved at søge ind på studier med højere adgangskvotienter. Model 4 viser desuden en lille effekt af uddannelsesstørrelse på kvindeandelen – mænd har en omend svag tendens til at søge ind på større studier.

Model 3 og 4 giver kun lidt ekstra indsigt i forhold til Model 1 og 2 – fra Model 2 til 3 falder deviance (Tabel A1), som er et mål for goodness-of-fit, kun marginalt. Med den tilgængelige data er Model 2 derfor et godt kompromis mellem antallet af variable og modellernes forklaringskraft. Det er klart at forventet indkomst samt uddannelsesretning er variablene med den primære forklaringskraft og kausale effekter bag kvindeandelen på lange videregående uddannelser.

Machine learning

I forrige afsnit var formålet at opstille en model, som kan give indsigt i kausale sammenhænge mellem kvindeandelen af de optagne studerende og de valgte forklarende variable. I kontrast til det, er formålet med dette afsnit at opstille en model, som kan forudsige kvindeandelen af de optagne studerende på baggrund af udvalgte inputvariable.

Ved brug af en *machine learning* algoritme udarbejdes et regressionstræ. Denne type regressionstræ benytter en delmængde af de tilgængelige data, kaldet træningssættet, til at klassificere studierne ved at "lære" forskelle i strukturer mellem uddannelser, og dermed forudsige udfaldet af kvindeandel. Metoden kaldes *supervised learning*, da algoritmen udregner og dernæst viser, hvilke karakteristika, der bestemmer divergerende niveauer af kvindeandele på uddannelser i træningssættet. I

dette tilfælde benyttes inputvariablene forventet gennemsnitsindkomst, uddannelsesretning, adgangskvotient samt uddannelsesstørrelse i et forsøg på at forudsige kvindeandelen af de optagene studerende på en given uddannelse.

Intuitionen bag regressionstræet er, at træningsdata gentagne gange opdeles i mindre dele. De enkelte knuder i træet illustrerer denne opdelingsproces. Til de enkelte knuder hører en mindre model, som igen splitter data op i mindre dele ved at fremsætte nye "spørgsmål", der besvares binært. Processen kaldes segmentering og er relativ simpel ift. andre algoritmer, dermed kan segmenteringsmodellen fortolkes intuitivt og nemt illustreres grafisk.

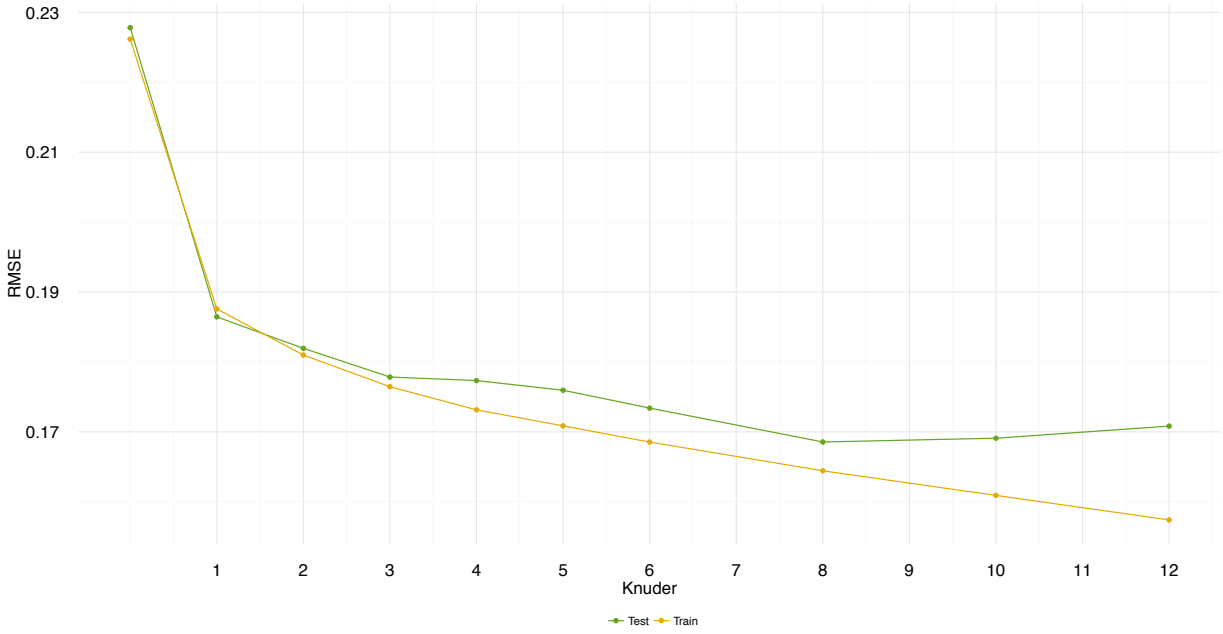
Det muligt at udarbejde et beslutningstræ med lige så mange knuder, som der findes observationer i træningsdatasættet og derved lade kvadratroden af kvardatafvigelsesgennemsnittet (RMSE) konvergere mod 0. Mange knuder gør modellens beskrivelse af træningsdata meget nøjagtig, men de gør samtidig modellen ude af stand til at give præcise forudsigelser på nye data på grund af overfit. Derfor er valget af et optimalt antal knuder centralt (Friedman et a.l. 2001).

Validering

Der findes forskellige metoder til udvælgelsen af antallet knuder, heriblandt krydsvalideringsmetoden, som ofte giver mere robuste resultater end simpel validering. På trods af dette, samt at krydsvalidering er standardmetoden i algoritmepakken, benyttes her simpel validering. Årsagen til dette er, at simpel validering giver mulighed for grafisk at forstå udvælgelsen af knuder.

Indledningsvist udtrækkes halvdelen af observationerne, på tilfældigvis, i det oprindelige datasæt. Udtrækket bliver brugt som træningsdata, mens den resterende del af data benyttes til validering, også kaldes testdata. Dernæst udarbejdes et stort regressionstræ, som beskæres indtil det optimale undertræ opnås. Det optimale undertræ er det undertræ, som mindsker RMSE mellem forudsigelserne og observationerne i testdata. Ved denne proces frembringes i dette tilfælde et regressionstræ, som består af 8 knuder. Dette fremgår af grafen nedenfor, hvor RMSE er en funktion af knuder. For testdata har RMSE globalt minimum i 8, mens RMSE for træningsdata konvergere mod 0, præcis som ventet. Det udtrykker altså balancen mellem bias ved lille et træ og overfitting ved et stort træ.

Figur 5: Kvadratroden af kvardatafvigelsesgennemsnittet for træning- og testdata



Kilder: Egne beregninger, UFM.dk og CEPOS

Den alternative valideringsmetode, krydsvalidering, minder meget om ovenstående simple valideringsmetode - dog betyder krydsvalidering at træningsdata opdeles i flere undersæt, og derved skal flere modeller estimeres og beskæres. Denne process er beregningstung, og derfor opstilles en cost-funktion for RMSE, hvori et kompleksitetsled formuleres. Kompleksitetsleddet er produktet af en positiv strafkoefficient og antal af knuder. Givet en række udvalgte strafkoefficienter estimeres blot en sekvens af de ellers mange kombinationer af modeller. Gennem en rekursiv krydsningsprocess vælges efterfølgende den strafkoefficient, der minimerer cost-funktionen for modellen i træningssættet. En fordel ved krydsvalidering er en lavere risiko for stikprøve-bias, da træningsdata opsplittes i flere tilfældige undersæt. Udover krydsvalidering kunne bootstrapping også benyttes til at lave syntetiske stikprøver. Den grafiske fortolkning bag model- eller knudevalg - som vist i Figur 5 - er dog kun mulig ved simpel validering. Således estmimeres regressionstræet med 8 knuder, som beskrives og illusteres i næste afsnit.

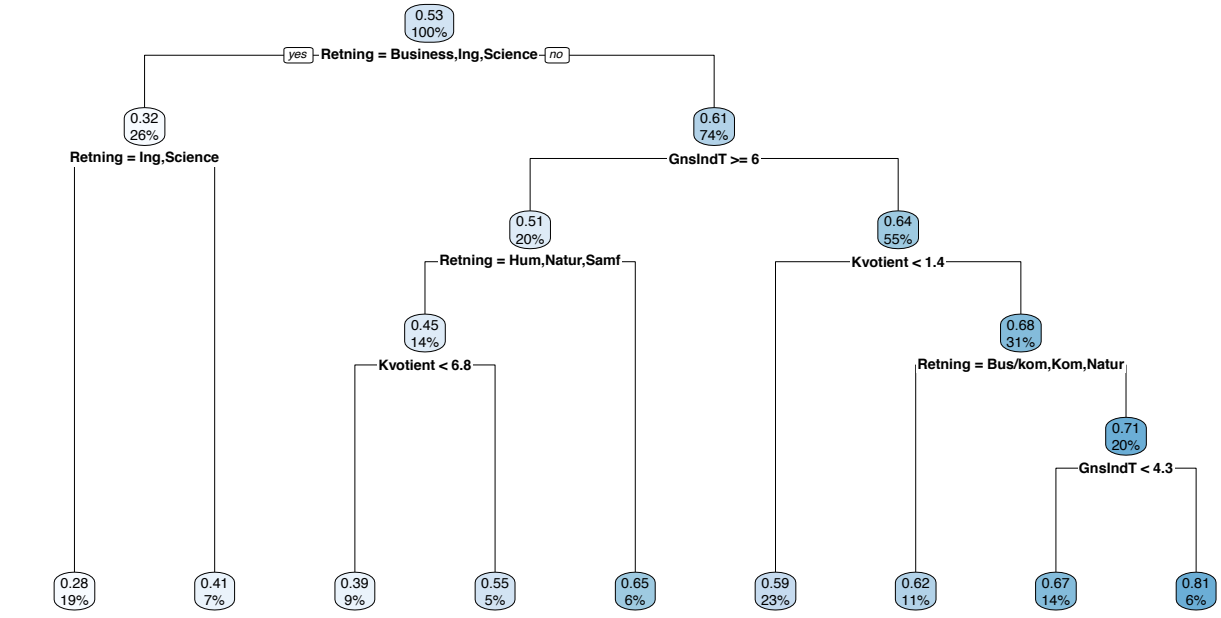
Regressionstræ

Ved den første knude i træet tages udgangspunkt i hele datasættet. Det ses her, at modellen på baggrund af alle data beregner et gennemsnit for kvindeandelen af optagne til 0,53. Modellen

fremsætter i første knude spørgsmålet *"Retning = business, ingeniør eller science"*, og data opdeles dernæst efter svaret på dette spørgsmål. Efterfølgende beregnes en ny værdi for den forventede kvindeandel af de optagne studerende med de nye oplysninger om uddannelsesretninger under de to undertræer in mente. Allerede ved dette skridt i træet er modellens forudsigelser interessante. Det fremgår, at i tilfældet, hvor retningen for uddannelserne er business, ingeniør eller science, vil den forventede kvindeandel af de optagne studerende falde fra 0,53 til 0,32. Er der derimod tale om de resterende uddannelsesretninger, såsom samfundsvidenskab, sundhed eller humaniora, vil den forventede kvindeandel stige til 0,61. Ved at indskrænke fokus til bestemte uddannelsesretninger ses dermed store variationer i fordelingen af mænd og kvinder i uddannelsessystemet. Ud fra de 26 % af data, hvor retningen er business, ingeniør eller science, vil modellen forudsige kvindeandelen af de optagne studerende til at være 0,41, hvis retningen er business, og 0,28, hvis den er ingeniør eller science. Dermed ser vi, at retningerne ingeniør og science umiddelbart er nogle af de uddannelsesretninger med meget lave kvindeandele.

I den modsatte side af træet ses vigtigheden af forventede lønninger. Kvindeandelen falder fra 0.61 til 0.51 på uddannelser med forventede lønninger over 600.000 kroner, og stiger fra 0.61 til 0.64 på uddannelser med forventede lønninger under 600.000 kroner. Efter spørgsmålet om forventet lønniveau fokuseres på uddannelsesretninger og adgangskvotienter i træet. Her falder kvindeandele for uddannelser der har med lavere adgangskvotienter.

Figur 6: Regressionstræ



Kilder: Egne beregninger, UFM.dk og CEPOS

Regressionstræet bekræfter dermed resultaterne fra logit-modellerne; inputvariablene uddannelsesretning og gennemsnitsindkomst har den største betydning for variation i kønsfordeling på de forskellige uddannelser.

Konklusion

De undersøgte data viser tydeligt, at uddannelsesvalg har stor betydning for fremtidige lønninger. Men lønninger har samtidig også stor betydning for uddannelsesvalg. Mænd vælger, aktivt eller passivt, uddannelser, der leder til bedre lønnet arbejde. Dermed kan kønsfordelingen på lange videregående uddannelser forklare en stor del af lønforskelle imellem kønnene, som observeres på arbejdsmarkedet. Kønsforskelle i uddannelsesvalg afhænger i høj grad af uddannelsesretning, men selv efter justering for uddannelsesretning vælger kvinder uddannelser, der leder til lavere lønninger sammenlignet med mænd.

Dermed starter kønsuligheden observeret på arbejdsmarkedet allerede, når studenter vælger deres bacheloruddannelse. Denne konklusion betyder, at policy målrettet kønsulighed skal starte allerede

her: Mindre kønsforskelle i uddannelsesvalg giver mindre ulighed på arbejdsmarkedet. Policy bør derfor takle socialt determinerede kønspræferencer i forhold til uddannelse, da disse på nuværende tidspunkt cementerer lønforskelle kønnene imellem på trods af samme resulterende niveau af human kapital. På arbejdsmarkedet bør policy fokusere på at begrænse negative konsekvenser af feminisering, i.e. lavere lønninger i faggrupper domineret af kvinder – dette er specielt vigtig i professionerne, hvor kvinder nu udgør størstedelen af optagne studerende.

Analysen af sammenhænge mellem uddannelsesvalg og lønforskelle mellem kønnene kunne gøres mere robust ved at benytte panel data og kønsopdelt løndata. På trods af disse datamangler giver analysen her en grundig, præliminær indsigt i sammenhængen, som ikke illustreres i standard human kapital-modeller af arbejdsmarkedet. Desuden kan analysen udbredes til alle uddannelser, ikke kun lange videregående uddannelser.

Litteraturliste

Baum, C.F., 2008. Modelling proportions. *Stata Journal,* 8(2), pp.299–303.

Becker, G., 1964. *Human capital; a theoretical and empirical analysis, with special reference to education,* New York: National Bureau of Economic Research; distributed by Columbia University Press.

Becker, G.S., 1985. Human Capital, Effort, and the Sexual Division of Labor. *Journal of Labour Economics,* 3(1), pp.S33–S58.

Becker, G.S., 1991. *Treatise on the family,* Enl. ed., Cambridge, MA: Harvard University Press.

Bergmann, B.R., 1974. Occupational Segregation, Wages and Profits When Employers Discriminate by Race or Sex. *Eastern Economic Journal,* 1(2), pp.103–110.

Blackburn, R.M. et al., 2002. Explaining gender segregation. *British Journal of Sociology,* 53(4), pp.513–36.

Connolly, S. og Gregory, M., 2007. Women and Work since 1970. I N. Crafts, I. Gazeley, & A. Newell, eds. *Work and Pay in 20th Century Britain.* Oxford: Oxford University Press, pp. 142–167.

Crompton, R., 2007. Gender inequality and the gendered division of labour. I J. Browne, ed. *The Future of Gender.* Cambridge: Cambridge University Press, pp. 228–249.

Friedman, J., Hastie, T. og Tibshirani, R. 2001. *Introduction to statistical learning. Vol. 1.,* Berlin: Springer Series in Statistics.

Hakim, C., 2000. *Work-lifestyle choices in the 21st century: preference theory,* Oxford: Oxford University Press.

Rubery, J., 2015. Change at work: feminisation, flexibilisation, fragmentation and financialisation. *Employee Relations,* 37(6), pp.633–644.

Appendix

Tabel A1: Logit koefficienter

Model	(1)	(2)	(3)	(4)
Konstant	0.955***	1.405***	1.088***	1.102***
	(0.020)	(0.038)	(0.040)	(0.040)
Indkomst i 100.000 kr.	-0.140***	-0.168***	-0.162***	-0.158***
	(0.003)	(0.005)	(0.005)	(0.005)
Business		-0.700***	-0.689***	-0.632***
		(0.033)	(0.033)	(0.034)
Humaniora		-0.167***	-0.074*	-0.065*
		(0.030)	(0.031)	(0.031)
Ingenior		-1.168***	-1.062***	-1.086***
		(0.036)	(0.036)	(0.036)
Jura		0.439***	0.307***	0.446***
		(0.042)	(0.042)	(0.045)
Kommunikation		-0.342***	-0.262***	-0.269***
		(0.034)	(0.034)	(0.034)
Naturvidenskab		-0.229***	-0.138***	-0.145***
		(0.032)	(0.032)	(0.032)
Samfundsvidenskab		-0.066*	-0.076*	-0.046
		(0.031)	(0.031)	(0.032)
Science		-1.184***	-1.061***	-1.061***
		(0.037)	(0.037)	(0.037)
Sundhedsvidenskab		0.738***	0.572***	0.633***
		(0.040)	(0.040)	(0.041)
Adgangskvotient			0.043***	0.045***
			(0.002)	(0.002)
Totalt optag i hundrede				-0.035***
				(0.004)

Model	(1)	(2)	(3)	(4)
McFadden R-sq.	0.1	0.4	0.4	0.4
Deviance	19854.7	12942.2	12271.8	12206.5

Kilder: Egne beregninger, UFM.dk og CEPOS. Standardfejl i parenteser.