

# Eksamens projekt

## My Table of content

- Section 1
- Section 2

## Section 1

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 forudsatte cirkularitet, er der gennem de seneste årtier sket dramatiske kønsforandringer i human kapital investeringer: kvinder udgør nu majoriteten af studerende på længere videregående uddannelser (Kilde). 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 længerevarende videregående uddannelser, hvor mænd i 2014 havde en bruttoindkomst 36,54 % højere end kvinder, jf. tabellen nedenfor.

Table 1: Bruttoløn fordelt på køn, 2014

	Mænd	Kvinder
Lange videregående uddannelser	670.133	490.782

Kilde: Danmarks Statistik

En mulig forklaring for disse lønforskelle er divergerende afkast på investeringer i forskellige længerevarende uddannelser. Uddannelsesvalg er præget af segregering: kun få videregående uddannelser har et optag af studerende med lige andele af mænd af 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 længerevarende videregående uddannelser, der leder til relativt lavere lønninger, og en tendens blandt mænd til at

søge mod længerevarende 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).

TJEK: Kønsoptdeling indenfor specifikke fagområder, e.g. tekniske uddannelser: søger mænd også her mod bedre betalte uddannelser?

Med kønsopdelt data for ansøgerantal samt optagne studerende på længere videregående uddannelse er det muligt at bestemme udstrækningen af kønssegregering i uddannelsesvalg. Ydermere, gør løndata for færdige kandidater det muligt at bestemme, hvorvidt der forefindes en sammenhæng mellem kønssegregering blandt optagne studerende og forventet løn efter afsluttet uddannelse. Sluttelig analyseres det, hvorvidt forventet lønniveau kan forudsige kønsfordelingen blandt ansøgere på længere videregående uddannelser.

## **Data, metode og etik**

Data er let offentligt tilgængelig og ligger åbnet på UFM's hjemmeside, derved har der ikke været rettighedsproblemer. Dog kan det problematiseres, at nogle uddannelser har meget få optagne, således at man vil kunne finde frem til de studerendes identitet vha. universiteternes biblioteker med BA opgaver eller muligvist med LinkedIn

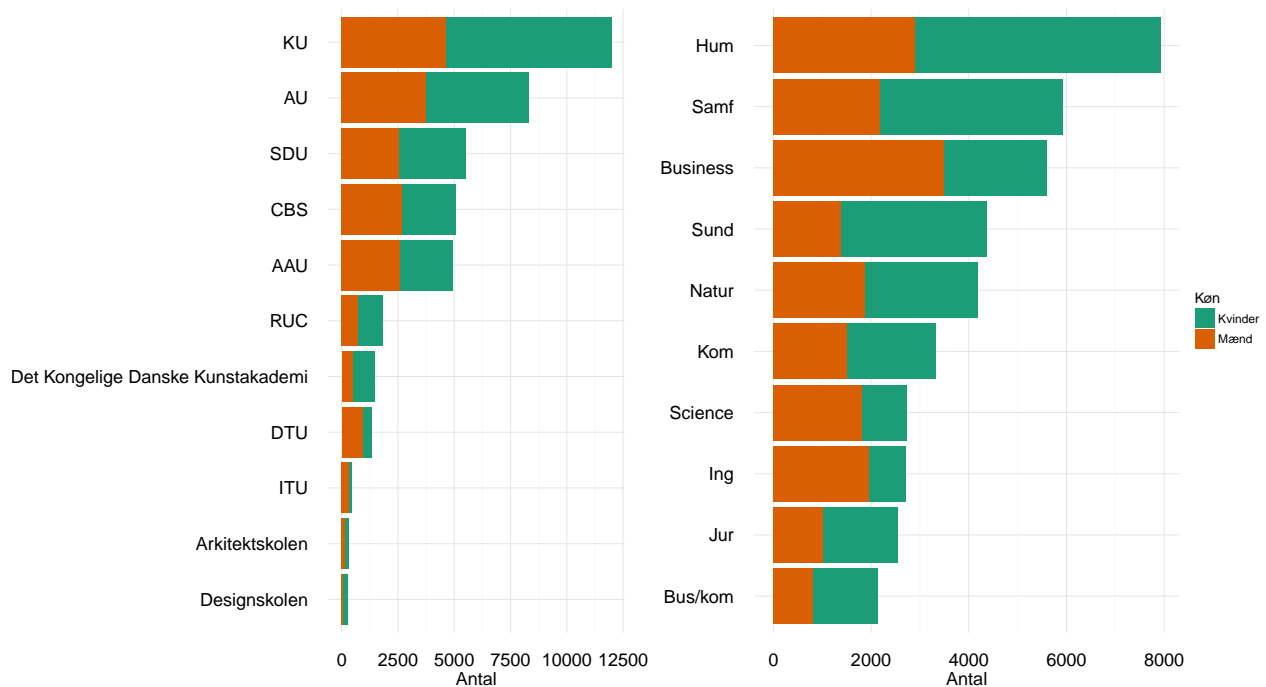
Metode Skrive her om hvilke datasæt vi har hentet og hvilke kilder: For hvert år har vi hentet: 1) ansøgninger fordelt på alle videregående uddannelser 2) optagne fordelt på alle videregående uddannelser 3) Adgangskvotienter for alle videregående uddannelser 4\*3

Vi har hentet CEPOS undersøgelsen for bruttoløn opgjort efter uddannelse.

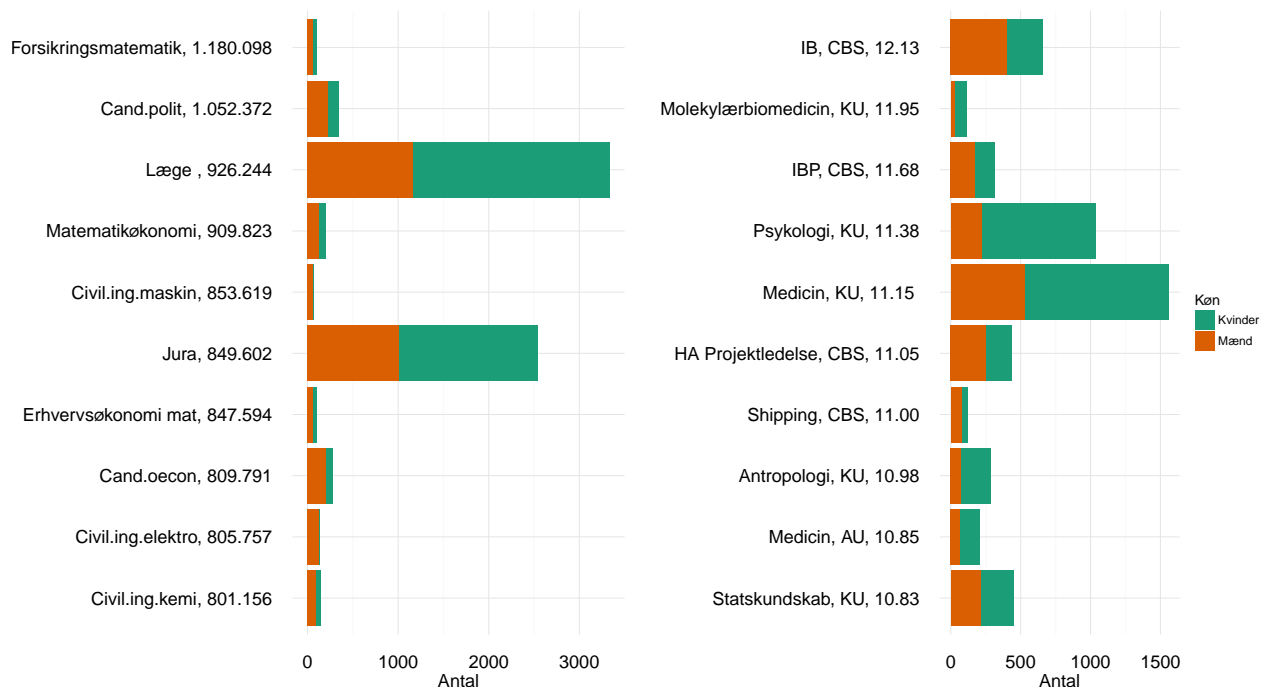
Dernæst gennemgik vi et datasæt kvalitativt henholdvis, at finde matchet mellem BA og færdig uddannelse og give hver uddannelse en retningskategori.

Vi joiner alle uddannelserne sammen på deres unikke uddannelsesnr. Dernæst joinet lønnen sammen fra CEPOS undersøgelsen ved benytte placeringsnr som nøgle.

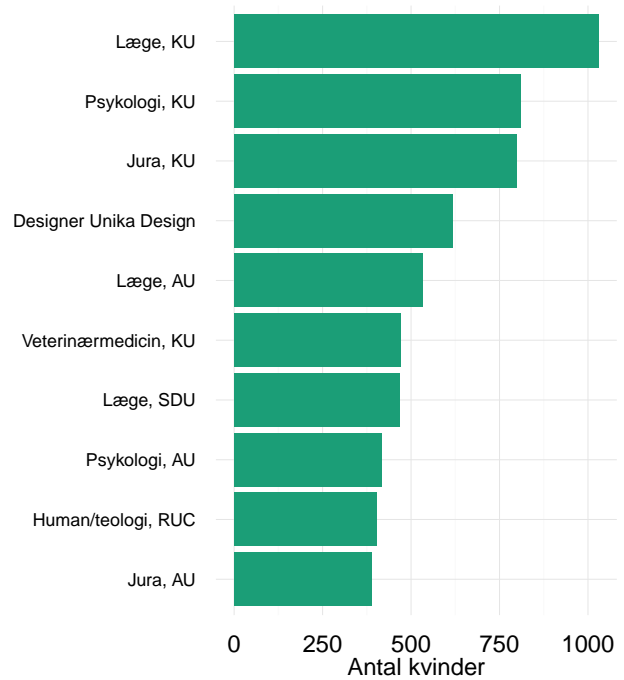
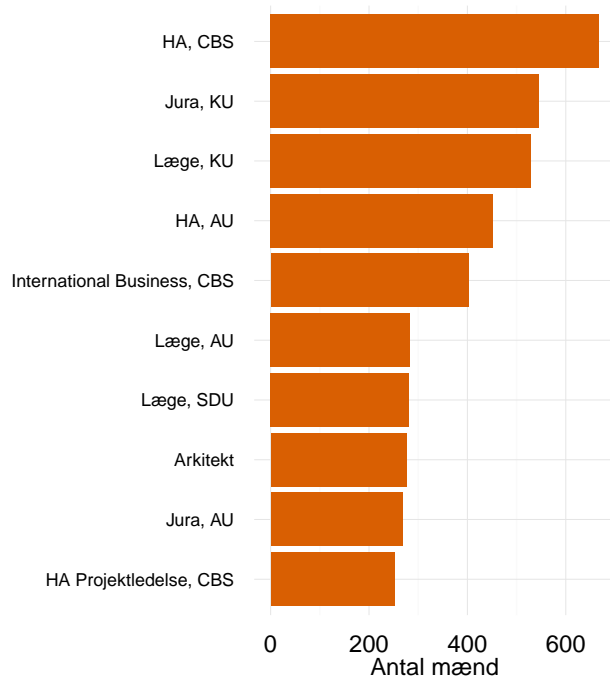
**Gennemsnitlig antal ansøgere fordelt på køn ved danske uddannelse institutioner og retninger, 2013-16**



**Gennemsnitlige antal ansøgere fordelt på køn for top 10 uddannelser målt på bruttoløn, kr. og top 10 gennemsnitlig højst kvotient, 2013-16**



## Top 10 gennemsnitlig mest søgte uddannelse for kønnene, 2013-2016



Tabel 2: Logit-modeller, marginal effekter

Model	(1)	(2)	(3)	(4)
Konstant	0.238*** (0.005)	0.350*** (0.010)	0.271*** (0.010)	0.274*** (0.010)
Indkomst i 100.000 kr.	-0.035*** (0.001)	-0.042*** (0.001)	-0.040*** (0.001)	-0.039*** (0.001)
Business		-0.174*** (0.008)	-0.172*** (0.008)	-0.157*** (0.008)
Humaniora		-0.042*** (0.008)	-0.018* (0.008)	-0.016* (0.008)
Ingeniør		-0.291*** (0.009)	-0.264*** (0.009)	-0.270*** (0.009)
Jura		0.109*** (0.010)	0.077*** (0.010)	0.111*** (0.011)
Kommunikation		-0.085*** (0.008)	-0.065*** (0.009)	-0.067*** (0.009)
Naturvidenskab		-0.057*** (0.008)	-0.034*** (0.008)	-0.036*** (0.008)
Samfundsvidenskab		-0.016* (0.008)	-0.019* (0.008)	-0.011 (0.008)
Science		-0.295*** (0.009)	-0.264*** (0.009)	-0.264*** (0.009)
Sundhedsvidenskab		0.184*** (0.010)	0.142*** (0.010)	0.158*** (0.010)
Adgangskvotient			0.011*** (0.000)	0.011*** (0.000)
Totalt optag i hundrede				-0.009*** (0.001)
McFadden R-sq.	0.1	0.4	0.4	0.4
Deviance	19854.7	12942.2	12271.8	12206.5

## Machine learning: Regressionsstræer

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

Ved brug af en *machine learning* algoritme udarbejdes et regressionstræ. Denne type regressionstræ vil på baggrund af en delmængde af den tilgængelige data, kaldet træningssættet, forsøge at klassificere studierne, ved at “lære” hvordan strukturen mellem forskellige studier er, og dermed forudsige udfaldet af kvindeandel. Metoden kaldes *supervised learning*, hvilket betyder at algoritmen for vist hvilke karakteristika der hører til forskellige niveau af kvindeandelene på studierne i træningsættet. I dette tilfælde benyttes variablene forventede gennemsnitsindkomst, studieretning, adgangskvotient og studiet størrelse som inputvariable til at forsøge at forudsige kvindeandelen af de optagede studerende på et givent studie.

Intuitionen bag regressionstræet er, at træningsdata gentagne gange opdeles i mindre dele. De enkelte knuder i træet illustrerer denne opdelingsproces idet der til de enkelte knuder hører en mindre model, som netop splitter data op i mindre dele ved at fremsætte nye “spørgsmål”, der skal besvares binært. Denne proces kaldes segmentering, og er relativ simpel ift. andre algoritmer. Segmentering giver intuitiv fortolkning af modellen, og kan nemt fremstilles grafisk. Det er naturligvis muligt at udarbejde et beslutningstræ med lige så mange knuder, som der findes observationer i træningsdata og derved lade RMSE konvergere mod 0. Dette vil gøre modellens beskrivelse af træningsdata meget nøjagtigt, men vil dog gøre modellen ude af stand til at give præcise forudsigelser på ny data, da modellen ville være overfittet. Derfor er det vigtigt, at vælge et optimalt antal knuder. Der findes forskellige metoder til udvælgelsen af knuder, heriblandt kryds-valideringsmetoden, som ofte giver mere robuste resultater end den mere simple valideringsmetode. På trods af dette, og at kryds-validering er standardmetoden i algoritmepakken, benyttes her den mere simple validering. Årsagen til dette er valideringsmetoden giver mulighed for grafisk, at forstå udvælgelsen af knuder.

## Validering

Indledningsvist udtrækkes halvdelen af observationerne i det oprindelige data-sæt, hvilke blev brugt som training set og den resterende del af data fungerede således som validation set. Dette betyder at beslutningstræet udarbejdes på baggrund af træningsdatasættet og derefter testes på testdata, således at vi kan vurdere modellen. cross validation og benyttes

Det kan dog ske at træet passer for godt til det datasæt der ligger til grund for dette, hvormed det risikeres at modellens forudsigelser ikke vil være gode. (<http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf>, s. 307-311)

Der findes

Vi ser altså, at der ved den første knude i træet tages udgangspunkt i hele datasættet. Vi kan herfra se, at modellen på baggrund af alle data for de tre inputvariable beregner et gennemsnit for kvindeandelen af optagede til 0,52. Modellen fremsætter i første knude spørgsmålet "Retning = business, ingeniør eller science" og data opdeles dermed efter svaret på dette. Ydermere beregnes en ny værdi for den forventede kvindeandel af de optagede studerende, givet de nye oplysninger om hvilke retninger der fokuseres på. Det er allerede ved dette skridt i træet interessant at se modellens forudsigelser. Vi ser, at i tilfældet hvor retningen for uddannelserne er business, ingeniør eller science vil den forventede værdi af kvindeandelen af de optagede studerende falde fra 0,52 til 0,32. Er der derimod tale om de resterende retninger for uddannelserne som eksempelvis samfundsvidenskab, sundhed eller humaniora vil den forventede kvindeandel stige til 0,6. Dette viser, at vi ved at indskrænke fokus til bestemte retninger inden for uddannelsesudvalget kan se en forskel i fordelingen af mænd og kvinder i uddannelsessystemet. Ser vi nu på de 28 % af data, hvor retningen er business, ingeniør eller science, vil modellen forudsige kvindeandelen af de optagede studerende til at være 0,4 hvis retningen er business og 0,29 for ingeniør eller science. Dermed ser vi, at retningerne ingeniør og science umiddelbart er nogle af de uddannelseskategorier, der ikke tiltrækker mange kvinder.

Det er interessant at se på knude...

Træet viser at det i beslutningsprocessen er inputvariablen retning der er den vigtigste, efterfulgt af gennemsnitsindkomst og afsluttende uddannelsesinstitution.

Kan bruges deskriptivt og



Hver knude viser - Den forudsagte andel af kvinder - Hvor procent af observationerne det gælder.

