

Avkastning av utdanning

En komparativ analyse av Norge og USA

Bacheloroppgave i Samfunnsøkonomi

Veileder: Bjarne Strøm

Mai 2019

Avkastning av utdanning

En komparativ analyse av Norge og USA

Bacheloroppgave i Samfunnsøkonomi
Veileder: Bjarne Strøm
Mai 2019

Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
Institutt for samfunnsøkonomi

Innhold

1. Innledning.....	2
1.1 Motivasjon.....	2
1.2 Problemstilling	2
2. Empirisk grunnlag	2
2.1 Mincer	2
2.2 Humankapital	3
2.3 Avkastning av utdanning.....	4
3. Økonometrisk teori.....	5
3.1 Innledning.....	5
3.2 Minste kvadraters metode (OLS)	5
3.3 Korrelasjon	6
3.4 Hypotesetesting	6
4. Presentasjon av data	7
4.1 Innledning.....	7
4.3 Deskriptiv statistikk.....	9
4.4 Kritikk av datasettet.....	12
4.5 Oppsummering	12
5. Regresjonsanalyse	12
5.1 Innledning.....	13
5.2 Valg av funksjonsform	13
5.3 Restriktiv modell	13
5.4 Utvidet restriktiv modell.....	14
5.5 Utvidet Mincer	16
5.5.1 Interaksjonsvariabelen <i>norge_yrs</i>	16
5.5.2 Alle kontrollvariabler	18
5.6 Heteroskedastisitet.....	21
5.7 Tolkning av resultat.....	22
6. Oppsummering og konklusjon	23
6.1 Oppsummering	23
6.2 Konklusjon	24
7. Referanser.....	24
Vedlegg	26

1. Innledning

1.1 Motivasjon

Økonomisk avkastning på utdanning varierer mellom land. Vi ønsker å undersøke hvordan utdanning påvirker lønnen i to land med svært ulik grad av fagorganisering og tariffavtaledekning. Vi har valgt å se på Norge og USA fordi 74% av Norges arbeidsstyrke er omfattet av tariffavtale, mens kun 13.3 % av arbeidsstyrken i USA er omfattet av samme type avtale (Hanushek, Guido, Wiederhold, & Ludger, 2015). Vi legger vekt på tariffavtaledekning fordi tariffavtale viser til en fagforeningsinnflytelse mange kan være omfattet av uten å personlig være medlem i en fagforening. Til tross for at både Norge og USA er OECD-land forventer vi å finne ulikheter i avkastning på utdanning, og disse ulikhetene ønsker vi å belyse på best mulig måte.

1.2 Problemstilling

Forventningene om ulikhetene vi nevner innledningsvis motiverte oss til å undersøke lønnsforskjellene mellom Norge og USA, og vi kom frem til følgende problemstilling;

«I hvilken grad kan vi observere forskjeller i lønnsavkastning av formell utdanning mellom land med ulik grad av fagorganisering?»

Vi ønsker også å undersøke i hvilken grad ulikheten i avkastning på formell utdanning påvirkes av arbeidserfaring, kognitive ferdigheter, kjønn, yrkesgruppe, sektor og stillingsprosent.

2. Empirisk grunnlag

2.1 Mincer

I denne oppgaven skal vi se blant annet se på effekten av utdanning på lønn. Mincers lønnsfunksjon, oppkalt etter økonomen Jacob Mincer, er en modell som forklarer inntekt som en funksjon av utdanning og arbeidserfaring. Vi antar at alle individer har likt utgangspunkt og like muligheter til å starte i et yrke, men yrkene varierer med tanke på hvilken grad av utdanning de krever. Tiden brukt på utdanning tilsvarer en utsettelse av inntekt til en senere alder. Yrker som krever høyere utdanning gir høyere lønn som kompensasjon for tiden brukt på utdanningen. Ulikhet i lønn mellom yrker er derfor en funksjon av ulikhet i utdanning (Mincer, 1958, s. 284).

I tillegg til skolegang er arbeidserfaring svært viktig for å prestere godt i alle yrker. Arbeidserfaring kan måles som antall timer brukt på arbeid. Økt erfaring og dermed økt produktivitet fører til lønnsøkning med alder opp til et visst punkt, hvor aldring fører til negativ

effekt på produktiviteten. En vesentlig forskjell mellom yrkesgruppene er at produktivitsveksten med alder er mer markant, og nedgang i produktivitet er mindre i yrker som krever lenger utdanning (Mincer, 1958, s. 287). Dagens standardmetode for å estimere avkastning på års utdanning er å estimere en log-lineær inntektsfunksjon basert på Mincers funn.

En svakhet ved Mincers lønnsfunksjon er at det er vanskelig å påvise kausalitet. De eneste kostnadene modellen tar hensyn til er alternativkostnadene ved å studere istedenfor å tjene penger ved arbeid. Modellen hadde vært sterkere dersom den hadde tatt hensyn til andre kostnader ved utdanning, som for eksempel prisen på utdanningen (Patrinos, 2016, s. 3). Korrelasjon kan være tilfeldig og er i utgangspunktet ikke et godt nok mål for å påvise årsakssammenheng (Thomas, 2005, s. 258).

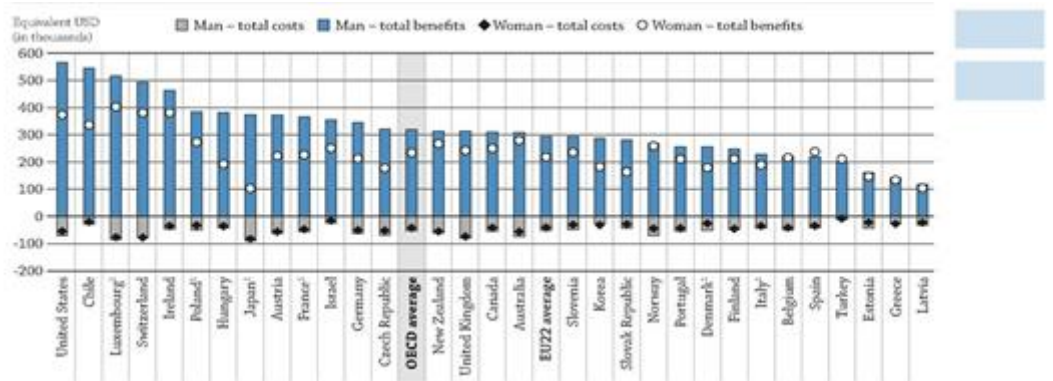
2.2 Humankapital

Gary Becker undersøkte avkastning på utdanning i 1964. Hans undersøkelse påpekte at formålet med utdanning er å øke individets fremtidige inntekt. Fra dette perspektivet blir utdanning sett på som en investering. Et individ må ha incentiver til å gjøre denne investeringen, og dette incentivet er at personen vil ende opp med en høyere lønn enn hva hen ville gjort uten utdanning (Hægeland, 2003, s. 198). Et utdanningsløp vil utelukke lønn underveis, og lønnen for en person med utdanning må derfor kompensere for dette. Investeringen gir ikke bare ren økonomisk avkastning; det reduserer risikoen for arbeidsledighet og øker personens sosioøkonomiske rang, i tillegg til de samfunnsøkonomiske gevinstene. Disse gevinstene er vanskelige å måle, men vi er under den generelle oppfatning om at gevinster som innovasjon og teknologisk kompetanse er blant de positive eksternalitetene ved høyere utdanning (Hægeland, 2003, s. 197).

Det antas at et individs høyere lønn ved høy utdanning også må ha en sammenheng med et økt produktivitsnivå - ellers ville arbeidsgivere tapt på å betale mer for lik effektivitet. Et omstridt tema er hvorvidt økte lønninger faktisk er et resultat av økt produktivitet, eller om det er filtrering av arbeidstakers attributter som ligger til grunn. Personers forutsetninger for å ta høyere utdanning er også omdiskutert – begrepet “ability bias” viser til at enkelte er mer akademisk anlagte enn andre, og vil derfor også ha større utbytte av å studere. Dette øker da sjansen for å få en godt betalt jobb etter fullført utdanning (Patrinos, 2016, s. 8).

2.3 Avkastning av utdanning

I OECD- rapporten “Education at a Glance” av 2018 (OECD, 2018) kommer det frem at det er en relativt liten gevinst av å ta høyere utdanning i Norge sammenliknet med andre OECD-land. USA topper statistikken som landet med høyest avkastning for formell utdanning.



- *Figur 2.3.1: Fremstilling av avkastning for utdanning i OECD- land. Bilde hentet fra rapport “Education at Glance”, 2018.*

Et lands system for lønnsfastsettelse kan påvirke størrelsen på lønnsforskjellene (Hægeland, 2003, s. 206). I Norge er lønnsforhandlingene sterkt sentraliserte og skjer stort sett gjennom fagforeninger. Internasjonale sammenligninger tyder på at sentraliserte lønnsforhandlinger medfører mindre lønnsforskjeller sammenliknet med de tilsvarende produktivitetsforskjellene. Samtidig er høy grad av sentralisering og lønnsammenpressing en grunn til å anta at lønnsstrukturen blir mindre fleksibel og dermed vil føre til lavere avkastning av høyere utdanning enn avkastningen av produktivitet (Hægeland, 2003, s. 206). Hægeland viser til statistikk fra OECD som påpeker at forskjellen i forventet arbeidsledighetsrate mellom grupper med ulikt utdanningsnivå i Norge er lavere enn gjennomsnittet i OECD, og mye mindre enn i USA. Forskjellene i tariffavtaledekning i Norge og USA kan være en årsak til at vi ser denne store forskjellen i avkastning av utdanning.

2.4 Oppsummering

Basert på empirien ønsker vi å undersøke om sektor, bransjeforskjeller, kjønn, ferdighetsnivå og stillingsprosent har innvirkning på avkastning av formell utdanning i Norge og USA, samt forskjellen mellom disse. Empirien tyder på at forskjellene i avkastning er til stede. Dette kan ha med strukturelle motsetninger å gjøre, som ulik grad av fagorganisering. Vi ønsker å undersøke om dette lar seg bevise økonometrisk videre i vår analyse.

3. Økonometrisk teori

3.1 Innledning

I dette kapittelet forklarer vi vår økonometriske metode, minste kvadraters metode, og hvordan den benyttes i regresjonsanalysen av datamaterialet.

3.2 Minste kvadraters metode (OLS)

I denne oppgaven skal vi analysere en eventuell årsakssammenheng mellom lønn og utdanning. OLS er en metode for å estimere de ukjente parameterne i en lineær regresjonsmodell. (Thomas, 2005, ss. 266-267). En forutsetning for å utføre en regresjonsanalyse er antakelsen om lineær årsakssammenheng mellom variabler. Disse variablene er henholdsvis den avhengige utfallsvariabelen Y og en eller flere uavhengige variabler, X_i . Den faktiske verdien på Y er ikke nødvendigvis lik forventningsverdien. Differansen mellom forventningsverdien $E(Y)$ og faktisk verdi på Y er gitt ved et støyledd, ε , som dekker de uforklarte og ofte umålbare faktorene som spiller inn på den forventede verdien til populasjonen.

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_i X_i + \varepsilon \quad (3.1)$$

Bruk av OLS krever fravær av eksakt sammenheng mellom x-ene, samtidig forlanger metoden sterke forutsetninger om restleddet. Dersom disse betingelsene ikke er oppfylt vil metoden kunne gi misvisende resultater. Forutsetningene for feilleddet er som følger (Thomas, 2005, ss. 356-359).

- (i) $E(\varepsilon_i) = 0$, restleddet har forventning lik null. Dette beskriver en situasjon der avviket mellom observerte verdier og estimerte verdier er lik null.
- (ii) $\text{Var}(\varepsilon_i) = \sigma^2 < \infty$: Feilleddets varians er konstant og uendelig for alle verdier.
- (iii) $\text{Cov}(\varepsilon_i), (\varepsilon_j) = 0$: Feilleddene er uavhengige av hverandre.
- (iv) $\varepsilon_i \sim N(0, \sigma^2)$: Feilleddet er normalfordelt, noe som er nødvendig for hypotesetesting.

OLS utføres ved å minimere summen av variansen, som er kvadratet til avviket mellom den estimerte og den observerte verdien. Dette gir oss estimater på konstantleddet (α) og stigningstallet (β) henholdsvis a og b , som gir oss en estimert regresjonslinje:

$$\hat{Y} = a + bX_i \quad (3.2)$$

OLS-estimatene gir oss den beste mulige linja, men vi må likevel beregne hvor godt regresjonslinja beskriver datasettet. Dette gjør vi ved å benytte føyningsmål eller determinasjonskoeffisienten R^2 (Thomas, 2005, s. 273) Denne viser hvor stor andel av total variasjonen i Y som kan forklares av variasjonen i X - altså andelen forklart variasjon (SSE) i forhold til total variasjon (SST) (Thomas, 2005, ss. 274-277).

$$R^2 = \frac{SSE}{SST} \left(= \frac{b^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} \right), 0 \leq R^2 \leq 1 \quad (3.3)$$

Siden føyningsmålet ligger mellom 0 og 1 vil forklaringskraften nesten alltid øke desto flere varianter av X som inkluderes i modellen. Dette gjør at R^2 kan gi misvisende mål. Benytter derfor en justert versjon av R^2 . Denne tar hensyn til at flere variabler tilføyes modellen og justerer forklaringskraften etter det (Thomas, 2005, s. 421).

3.3 Korrelasjon

Korrelasjonskoeffisienten ρ forteller oss hvor sterk samvariasjon det er mellom to eller flere variabler. Vi kan måle korrelasjonen i et utvalg og få et bilde av hvordan populasjonens korrelasjonskoeffisient ser ut. Utvalgets korrelasjonskoeffisient er definert som:

$$R = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}}, -1 \leq R \leq 1 \quad (3.4)$$

Er det en eksakt lineær sammenheng mellom X og Y, vil $R = +1$ v -1 . Dersom det ikke finnes noen lineær sammenheng mellom variablene er $R = 0$. Perfekt sammenheng mellom X-ene bryter med OLS-forutsetningene siden det blir umulig å estimere separate effekter av ulike X-er, samt å finne separate β -er, kun en estimert kombinasjon. (Thomas, 2005, ss. 257-258).

Det er viktig å skille mellom korrelasjon og kausalitet. Korrelasjon viser *samvariasjon*, ikke nødvendigvis sammenheng. Selv om to fenomener korrelerer betyr det ikke nødvendigvis at sammenhengen er kausal. Man kan oppleve det som kalles en spuriøs årsakssammenheng, som betegner en situasjon der det er en tredje variabel, Z, som påvirker X og Y slik at det ser ut som at det er en årsakssammenheng mellom X og Y (Thomas, 2005, s. 258).

3.4 Hypotesetesting

Hypotesetest er en metode som anvendes for å avgjøre om OLS-estimatorene stemmer overens med virkeligheten under en gitt prosentandel usikkerhet. Denne usikkerheten kalles signifikansnivået. Metoden krever en nullhypotese (H_0) og en alternativhypotese (H_A), hvor H_0 er hypotesen man vil undersøke om har forkastningsgrunnlag, mens H_A er en komplementær hypotese man ønsker å underbygge (Thomas, 2005, ss. 369-371). Det er testobservatoren (TS)

og kritisk verdi som avgjør hvilke verdier som kan forkaste H_0 . I statistiske tester er det vanlig å benytte et 5% signifikansnivå. Signifikansnivået uttrykker sannsynligheten for å forkaste en gyldig nullhypotese (Thomas, 2005, s. 129). Ut ifra hvilken test man gjennomfører vil man få ulik kritisk verdi og testobservator som avgjør om H_0 må forkastes eller ikke.

Man kan teste hypoteser om individuelle parametere, samt utføre multiple hypotesetester om samtlige parametere. t-testen undersøker om effekten av enkeltparametere på avhengig variabel er signifikante når populasjonsvariansen er ukjent (Thomas, 2005, s. 428). t-testen er en statistisk test som blant annet sammenligner estimerte koeffisienter i to datasett (Thomas, 2005, ss. 158-161). For å kunne teste multiple hypoteser på en gang og om de kollektivt har signifikant effekt på avhengig variabel må vi gjennomføre en F- (Thomas, 2005, ss. 415-419). Denne testen baserer seg på endring i residualvariasjonen (SSR) når restriksjoner pålegges en modell. (Thomas, 2005, ss. 439-443). Altså den relative endringen i forklaringskraften med og uten restriksjoner. Stor økning i residualvariasjonen tyder på ugyldig restriksjon og forkastelse av H_0 .

4. Presentasjon av data

4.1 Innledning

Datasettene som blir brukt er hentet fra PIAAC-undersøkelsen i regi av OECD som foregikk fra august 2011 til mars/april 2012, i 24 land. Målet med undersøkelsen var å gjøre en kartlegging av den voksne befolkningens ferdigheter i henholdsvis leseferdigheter, tallforståelse og problemløsning i IKT-miljø. Disse danner grunnlaget for videre ferdighetsutvikling og omtales som nøkkelferdigheter for å bearbeide informasjon. PIAAC er en del av flere internasjonale undersøkelser som forsøker å kartlegge ulike deler av befolkningens ferdigheter, og kan ses i sammenheng med resultatene fra de tidligere undersøkelsene International Adult Literacy Survey (IALS) og Adult Literacy and Life Skills (ALL). Det er også gjort et omfattende arbeid for at PIAACS resultater skal være sammenlignbare på tvers av landegrenser, både når det gjelder metode og innsamling (Bjørkeng, 2013). For å unngå seleksjonsproblem knyttet til de yngre som fortsatt er under utdanning, bruker vi kun resultatene rapportert inn fra de over 25 år.

4.2 Presentasjon av datasett

Datasettet består av en avhengig variabel, kontrollvariabler og dummyvariabler. Den avhengige variabelen *learnhrppp* beskriver logaritmen til timelønn oppgitt i dollar, kjøpekraftjustert. En

viktig kontrollvariabel er formelt utdanningsnivå, *yrsqual*, som er målt som høyeste utdanningsnivå oppnådd beregnet i antall år utdannelse. *c_q09_c* forteller hvor mange år respondenten har vært i lønnet arbeid og *c_q09_c_exp* tillater en ikke-lineær sammenheng av denne.

d_q03 forteller hvilken sektor respondenten jobber i. Den er lik 1 hvis det er privat sektor, 2 hvis offentlig og 3 dersom vedkommende jobber for en interesseorganisasjon. Sammen med variabelen for yrkesgruppe, *isco1c*, utgjør de såkalte indikatorvariabler som defineres ved ulike tallkoder. Videre har vi også variablene for de kognitive ferdighetene (*litscore1*, *numscore1*, *pslscore1*) hvor vi har valgt å bruke de standardiserte scorene fordi det gir sammenlignbare og benevningsfrie tolkninger av koeffisienter på tvers av måleskalaer og studier. *isco1c* forteller hvilken bransje eller yrkesgruppe respondenten tilhører. Den omfatter tallkodene 2-11 som benevner ulike kategorier, disse står oppført i tabell V1 i vedlegget. Alle variablene er listet i tabell 4.1, med en kort forklaring av hver enkelt variabel.

Vi konstruerer også egne dummer som kontrollerer for nasjonalitet og avkastningsforskjell mellom landene. En dummy-variabel er en binær variabel som benyttes for å måle forhold som ikke naturlig kan representeres kontinuerlig (Thomas, 2005, ss. 424-425). Dummyen *norge* viser forskjeller i lønnsnivå med USA som referanseland. *norge_yrs* viser avkastningsforskjellen på den formelle utdanningen, også her med USA som referansekategori. Har også dummyvariablene *female* og *full_time* som henholdsvis sier til kjønn og om vedkommende har fulltidsjobb.

Tabell 4.1 - Oversikt over variabler

<i>earnhrppp</i>	Timelønn oppgitt i USD. Kjøpekraftsjustert
<i>full_time</i>	Fulltidsjobb, lik 1 hvis respondenten er i fulltidsjobb
<i>c_q09_c</i>	Erfaring, definert som antall år med lønnet arbeid
<i>c_q09_exp</i>	Erfaring kvadrert
<i>female</i>	Kjønn, lik 1 hvis kvinne, lik 0 hvis mann
<i>yrsqual</i>	Formelt utdanningsnivå
<i>litscore1</i>	Kognitive ferdigheter målt ved lesetest, standardisert

<i>numscore1</i>	Kognitive ferdigheter målt ved regnetest, standardisert
<i>pslscore1</i>	Kognitive ferdigheter målt ved IKT-problemløsning, standardisert
<i>d_q03</i>	Om respondenten jobber i offentlig eller privat sektor, eller tilhører NGO
<i>isco1c</i>	Bransje/yrkesgruppe
<i>norge</i>	Nasjonalitet
<i>norge_yrs</i>	Avkastningsforskjell av formell utdanning
<i>learnhrppp</i>	Logaritmen til <i>earnhrppp</i>

4.3 Deskriptiv statistikk

Tabell 4.2 - Deskriptiv statistikk for avhengig variabel lønn

	Alle	Norge	USA
Gjennomsnitt	24.07909	25.51268	22.30152
Minimum	5.892016	5.892016	7.4
Maximum	64.50208	64.50208	51.5
Standardavvik	10.92847	8.597519	13.04644
Observasjoner	5461	3023	2438
Median	22	23.67661	18.2

Den deskriptive statistikken viser at det er forskjeller i lønningene i de to landene. Det gjennomsnittlige lønnsnivået er ca. \$3,2 høyere i Norge enn USA, noe som ikke er overraskende med tanke på at målt BNP per innbygger er høyere i Norge enn i USA (FN-Sambandet, 2017).

Selv om det er mindre forskjell mellom høyeste og laveste lønning i USA, ser vi på standardavvikene at den gjennomsnittlige spredningen i lønningene er lavere i Norge. Dette kan tyde på at Norge opplever mindre ulikheter i lønningene eller at en betydelig del av den arbeidsdyktige befolkningen får en lønning som er sentrert rundt den gjennomsnittlige lønningen, som ifølge Hægeland kan være et resultat av sentraliserte lønninger (Hægeland, 2003).

Det kan også være nyttig å se på medianen når man skal se på ulikheter, da denne kan gi et riktigere bilde av fordeling enn gjennomsnittet, som kan bli påvirket av ekstremverdier og dermed gi et noe uriktig bilde av fordelingen. Ser at det er forskjell på median- og gjennomsnittslønnen i begge landene, men det er en vesentlig større forskjell i USA, noe som igjen kan peke på at det er større ulikheter i lønnsnivået i USA enn i Norge.

Tabell 4.3 - Deskriptiv statistikk for Norge og USA

Variabler	Norge		USA	
	Gjennomsnitt	Standardavvik	Gjennomsnitt	Standardavvik
<i>yrsqual</i>	14,66	2,45	13,91	3,02
<i>C_q09_c</i>	21,69	11,61	22,73	12,2
<i>C_q09_c</i>	3.42e+18	2.66e+19	5.78e+18	3.42e+19
<i>female</i>	0,48	0,50	0,55	0,50
<i>full_time</i>	0,69	0,46	0,63	0,48
<i>litscore1</i>	0,03	1,01	0,0026	1,02
<i>numscore1</i>	0,04	1,01	0,0225	1,02
<i>palscore1</i>	-0,06	1,01	0,03	1,02
<i>D_q03</i>	1,41	0,52	1,37	0,62
<i>isco1c</i>	5,23	2,64	5,67	2,95

Ser at gjennomsnittlig arbeidserfaring er 21,69 år i Norge og 22,73 år i USA. Gjennomsnittlig utdanning er i underkant av et år mer i Norge, og vi ser på standardavviket at det også er noe mindre spredning i utdanningsnivået i Norge.

Tabell 4.4: Gjennomsnittslønn ved høyeste oppnådde utdanning i antall år

Høyeste oppnådde utdanning	Norge	USA
6		12.9925
7	16.72824	
9		13.59746
10	20.99459	
12	21.7322	16.33444

13	22.71958	19.32736
14	23.53137	21.88138
15	25.8024	
16	26.5826	26.91606
18	30.08447	31.49966
19		32.66765
21	32.53181	33.74889

Tabell 4.4 viser at fra år 6 til 15 har Norge høyere gjennomsnittslønn til gitt utdanningsnivå. Ved høyere utdanning enn 15 år har USA i gjennomsnitt høyere lønninger. Ut ifra denne ser vi at det foreligger forskjeller i avkastningen av utdanning.

Tabell 4.5: Lønnsnivå fordelt på de ulike yrkesgruppene (variabel isco1c)

Bransjeklassifisering	Norge	USA
2	31.33443	31.90988
3	28.2106	29.3972
4	27.34624	23.82807
5	21.57486	16.8098
6	20.23504	14.2298
7	22.30213	20.38182
8	24.51176	19.63911
9	23.81564	15.31961
10	18.33956	13.57187

Tabellen viser at yrkeskategori 2,3,4 har de høyeste gjennomsnittslønningene i begge landene, dette er typiske lederstillinger, og yrker som krever høy utdanning, f.eks. ingeniører og jurister. Deretter er det noen ulikheter, for eksempel innad i USA lønnes arbeidere innen landbruk og fiske relativt høyere enn hva de gjør i Norge. På bunn i begge landene finner vi servicemedarbeidere og elementære yrker. Ser at gjennomsnittslønnen i Norge jevnt over er høyere, bortsett fra i kategori 2 og 3. Videre ser vi også at USA har større spredning i lønnsnivået enn Norge. Det høyeste lønnsnivået er noenlunde likt, men ser at USA har mye lavere lønnsnivå i mange av kategoriene.

4.4 Kritikk av datasettet

I datasettet er lønna oppgitt i timelønn, noe som kan resultere i lavere mål på avkastningen av utdanning. Redusert avkastning kan komme av at individer med høyere utdanning ofte jobber flere timer og vil følgelig tjene mer. Avkastningen av utdanning vil gi større utslag på estimatene om lønna hadde blitt målt i ukentlig eller årlig inntekt (Patrinos, 2016, s. 3).

I PIAAC-undersøkelsen vi tar utgangspunkt i, er lønnen rapportert ulikt for Norge og USA. Norge har rapportert kontinuerlig lønn, mens USA har rapportert lønn i deciler. Det vil si at alle lønnsdataene er sortert i stigende rekkefølge og deretter inndelt i ti like store deler. Da vil de i utvalget med 10% høyest lønn være i decil 10 og de med 10% lavest lønn være i decil 1. Denne ulikheten i rapportering av lønn kunne vært en svakhet ved datasettet, men ifølge rapporten vi tar utgangspunkt i (Hanushek, Guido, Wiederhold, & Ludger, 2015, s. 108) viser det seg at dette ikke skal ha substansiell innvirkning på resultatene vi får. På en annen side vil det ikke være nyttig å illustrere lønnsforskjellene mellom USA og Norge i et histogram på grunn av decilmålingen.

Det hadde også vært interessant å undersøke effekten av andre variabler som ikke tilhører datasettet. Hadde datasettet inneholdt en variabel for kostnader ved utdanning, kunne vi trolig fått et mer presist mål på avkastningen av formell utdanning. Hadde vi hatt variabler for hvorvidt respondentene i undersøkelse er fagorganisert kunne vi sett på direkte effekt av fagorganiseringsgrad på avkastning av utdanning.

4.5 Oppsummering

Datasettet vårt består av 5461 observasjoner som danner et godt grunnlag for å utføre ulike analyser. Siden vi skal sammenligne to land har vi valgt å benytte samlet datasett til utførelse av disse analysene. For å undersøke avkastningen av utdanning har vi valgt å analysere hvordan erfaring, ferdigheter, kjønn, nasjonalitet, næring- og sektortilhørighet påvirker avkastning av formelt utdanningsnivå. For å forstå de antatte lønnsforskjellene i landene kunne vi ha inkludert flere variabler i datasettet vårt, men har valgt å ikke gjøre det for å unngå en reduksjon i sammenligningsgrunnlaget fra tidligere analyser, samt at vi ikke har nok tilgjengelig datamateriale til å lage fullstendig virkelighetsnære modeller.

5. Regresjonsanalyse

5.1 Innledning

Vi skal i denne delen av oppgaven analysere datamaterialet ved hjelp av minste kvadraters metode i Stata. Vi tar først for oss en enkel Mincer inntektsfunksjon, som vi trinnvis utvider med flere kontrollvariabler for å undersøke hvorvidt lønnsnivået og avkastningen av formell utdanning påvirkes av disse.

5.2 Valg av funksjonsform

En forutsetning ved bruk av minste kvadraters metode er at det er en lineær sammenheng mellom variablene. En lineær modell alene gjør det vanskelig å sammenligne to ulike datasett. Benytter derfor en eksponentiell funksjon som kan transformeres til lineær form ved hjelp av naturlige logaritmer. Da vi funksjonen se ut slik:

$$\ln(Y) = \alpha + \beta_1 X_i \quad (5.1)$$

Dette kalles en log-lineær funksjonsform. Log-lineære forklaringsmodeller gjør det mulig å sammenligne estimerte verdier på tvers av studier og land, siden estimert lønnsforskjell kan tolkes som det prosentvise lønnsgapet uavhengig av måleenhet og avhengig variabel. Denne egenskapen gjør at log-lineære modeller egner seg for sammenligning på tvers av datasett.

5.3 Restriktiv modell

Vi starter regresjonsanalysen med en svært restriktiv modell som hverken tar hensyn til at gjennomsnittlig timelønn eller avkastningen på utdanning kan være ulik i Norge og USA. Vi bruker Mincers inntektsfunksjon der *learnhrppp* er den avhengige variabelen, *Y*, som representerer timelønn i dollar justert for kjøpekraft. Modellen kontrollerer for hvordan formelt utdanningsnivå (*yrsqual*), erfaring (*c_q09_c*) og en ikke-lineær effekt av erfaring (*c_q09_c*²) påvirker lønnsnivået. Velger å starte med en slik modell på grunnlag av Mincer sin teori om hvordan utdanningsnivå og erfaring er avgjørende for lønn. I modellen benytter vi både en lineær og en eksponentiell variabel av erfaring.

$$\text{Modell 1: } \ln(\text{earnhrppp}) = \alpha + \beta_1 \text{yrsqual} + \beta_2 \text{c_q09_c} + \beta_3 (\text{c_q09_c})^2 + \varepsilon_i \quad (5.2)$$

Utfører en OLS-regresjonsanalyse i Stata for å estimere parameterne β_i og predikere *Y*. Ved OLS-analyse blir støyledet lik 0 (Thomas, 2005, s. 359). Får følgende estimerte parametre:

$$\ln(\text{earnhrppp}) = 1.727874 + 0.080041 \text{yrsqual} + 0.0088314 \text{c_q09_c} - 5.65e-22 (\text{c_q09_c})^2 \quad (5.3)$$

Parameterne i modell 1 representerer den restriktive modellen, og den viser at ett års lengre utdanning vil føre til en avkastning på omtrent 8%. Denne økningen er gjennomsnittlig for begge land, siden vi ikke har noen variabel som kontrollerer for land. I tillegg er avkastningen på arbeidserfaring positiv, men avtakende. Timelønnen øker med ca. 0,8% per år med ekstra arbeidserfaring, men grunnet erfaring kvadrert vil avkastning på timelønn avta etter et visst punkt. Dette samsvarer med Mincers intuisjon om den ikke-lineære effekten av arbeidserfaring. Vi ønsker å teste om parameterne er signifikante og om de har en innvirkning på lønna eller ikke. Til dette benytter vi en t-test med hypotesene:

$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0$$

Velger signifikansnivå på 5%, og har dermed at kritisk verdi er 1,96, funnet i t-fordelingstabellen i (Thomas, 2005, s. 587). Dersom parameterens absolutte t-verdier overstiger kritisk verdi forkaster vi nullhypotesen og konkluderer med at verdiene er signifikante til et 5% signifikansnivå.

Tabell 5. 1 – t-verdier

t-test	Utdanning	Erfaring	Erfaring ²
T-verdier hentet fra Stata	37,57	17,29	-2,59

Ser at alle t-verdiens absoluttverdi er større enn kritisk verdi. Dette gir et grunnlag for å si at lønn påvirkes av utdanning og erfaring, men den restriktive modellen skiller ikke mellom lønnsnivå og avkastningen mellom landene og kan dermed ikke fortelle oss om det er noen systematiske lønnsforskjeller. I tillegg ser vi at den justerte determinasjonskoeffisienten R^2 i denne inntektsfunksjonen er 0.2361, som vil si at utdanning og inntekt forklarer omtrent 23,61% av variasjonen i lønn.

5.4 Utvidet restriktiv modell

Tar nå hensyn til at lønnsnivået kan være ulikt i Norge og USA og utvider Mincers lønnsfunksjon ved å konstruere en dummy-variabel.

Ønsker først å kontrollere for ulike lønnsnivå i landene og gjør dette ved å innføre en additiv dummy (Thomas, 2005, s. 425). Kaller denne dummyen for *norge*. Denne vil endre

konstantleddet i modellen og tillate at nivået på lønningene kan være forskjellig mellom landene. Dummyen vil representere interessevariabelen land og få verdien 1 hvis observasjonen tilhører et norsk individ, og 0 hvis individet er amerikaner. Med en ny forklaringsvariabel vil modellen se ut som følger:

$$\text{Modell 2: } \ln(\text{learnhrppp}) = \alpha + \beta_1 \text{yrsqual} + \beta_2 c_q09_c + \beta_3 2 c_q09_c^2 + \gamma_1 \text{norge} + \varepsilon_i \quad (5.4)$$

Hvis individet man observerer er norsk vil koeffisienten og opprinnelig konstantledd adderes og vi får et nytt konstantledd i modellen vår. Den additive dummyen legger altså til en verdi (γ_1) på opprinnelig konstantledd (α) (Thomas, 2005, s. 425).

$$Y = (\alpha + \gamma_1) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3^2 + \varepsilon_i, \quad D_i=1 \quad (5.5)$$

Ved observasjon av en amerikaner vil dummyen få verdien 0 og koeffisienten til dummyvariabelen vil forsvinne fra likningen, γ_1 .

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3^2 + \varepsilon_i, \quad D_i=0 \quad (5.6)$$

Vi ser at konstantleddet for nordmenn og amerikanere blir ulikt ved bruk av den additive dummyen, slik at lønnsnivå mellom individer med ulik nasjonalitet blir ulik. Videre vil bruk av log-lin funksjoner gjøre at γ_1 angir den prosentvise lønnsdifferansen, altså relativ lønnsforskjell mellom nordmenn og amerikanere. Ser at koeffisientene fra modell 1 endres lite ved tilførsel av dummy.

Ser av regresjonen gjort av modell 2 at for gitt utdanning har nordmenn i snitt 18,8 % høyere timelønn enn amerikanere. Dette stemmer overens med observasjonen gjort i tabell 4.2, hvor man så at spredningen i lønnsutvalget i Norge var mindre enn i USA. Standardavvikene var henholdsvis 8,6 og 13,0.

Vi tester om lønnsnivået i Norge er signifikant høyere enn i USA ved å utføre en t-test av γ_1 . Fra Stata har vi at t-verdien til *norge* er 17,03. Vi benytter et 5% signifikansnivå og utfører en ensidig hypotesetest der nullhypotesen er at lønnsnivået er likt for Norge og USA, mens alternativhypotesen er at lønnsnivået i Norge er høyere enn i USA. Kritisk verdi blir 1,64 (Thomas, 2005, s. 587). Siden t-verdi er høyere enn kritisk verdi, forkaster vi H_0 og kan konkludere med at Norge har høyere gjennomsnittlig timelønn enn USA. Forklaringskraften til modellen har også steget fra 23,61% til 27,66%.

Tabell 5.2: Estimeringsresultater for modell 1 og 2

	(1)	(2)
VARIABLES	learnhrppp	learnhrppp
yrsqual	0.0800*** (0.00213)	0.0760*** (0.00209)
c_q09_c	0.00883*** (0.000511)	0.00898*** (0.000497)
c_q09_c^2	-5.65e-22 *** (2.18e-22)	-4.90e-22 ** (2.12e-22)
norge		0.188*** (0.0110)
Observations	5,161	5,161
Adjusted R-squared	0.2361	0.2766

Standardavvik i parentes

*** p<0.01, ** p<0.05, * p<0.1

5.5 Utvidet Mincer

5.5.1 Interaksjonsvariabelen *norge_yrs*

Har nå vist at det er signifikante lønnsnivåforskjeller i USA og Norge, men vi vet fortsatt ikke hvordan avkastingsdifferansen på utdanning ser ut mellom landene.

Vi legger til et interaksjonsledd i regresjonsanalysen, ved å multiplisere dummyen *norge* og antall år utdanning, *yrsqual*. Vår nye variabel *norge_yrs* viser at avkastning av formell utdanning i Norge avviker fra avkastningen på samme formelle utdanning i USA.

$$\ln(\text{learnhrppp}) = \alpha + \beta_1 \text{yrsqual} + \beta_2 c_q09_c + \beta_3 c_q09_c^2 + \gamma_1 \text{norge} + \gamma_2 \text{norge} * \text{yrsqual} + \varepsilon_i \quad (5.7)$$

For Norge, med $D_i=1$, vil ligningen se ut som følger:

$$\ln(\text{earnhrppp}) = (\alpha + \gamma_1) + (\beta_1 + \gamma_2) \text{yrsqual} + \beta_2 c_q09_c + \beta_3 c_q09_c^2 + \varepsilon_i \quad (5.8)$$

Den første parentesen viser til konstantleddet. Andre parentes representerer avkastningen av utdanning i Norge og hvor mange prosentpoeng denne avviker fra avkastningen i USA under samme formelle utdanningsnivå.

For USA vil dummyen bli lik null, $D_i=0$, noe som resulterer i:

$$\ln(\text{earnhrppp}) = \alpha + \beta_1 \text{yrsqual} + \beta_2 c_q09_c + \beta_3 c_q09_c^2 + \varepsilon_i \quad (5.9)$$

Siden *norge* settes lik null vil både γ_1 og γ_2 forsvinne og avkastningen på utdanning i USA vil bli β_1 . γ_2 blir også det eneste som skiller avkastning av utdanning i de to landene. Med andre ord vil estimert effekt av variabel *norge_yrs* på lønn avhenge av nivået på γ_2 .

Får følgende stigningstall til forklaringsvariablene i Stata:

$$\text{Modell 3: } \ln(\text{earnhrppp}) = 1,435 + 0,0933\text{yrsqual} + 0,00857c_q09_c - 5.41e - 22c_q09_c^2 + 0,739\text{norge} - 0,0377\text{norge_yrs} \quad (5.10)$$

Siden det er inkludert en interaksjonsvariabel mellom *norge* og *yrsqual* vil koeffisienten til *yrsqual* nå vise til avkastningen på utdanning for amerikanske individer. *yrsqual* viser at ett års ekstra utdanning i USA i gjennomsnitt gir 9,33% høyere timelønn. Samtidig viser koeffisienten til *norge_yrs*, γ_2 , at den samme ytterligere utdanningen i Norge gir 3,77 prosentpoeng lavere timelønn enn i USA, altså 5,56% økt timelønn etter ett år lenger utdanning i Norge. Dermed viser denne modellen at avkastningen av utdanning er større i USA enn i Norge. Kjører en t-test av for å undersøke om påstanden er signifikant:

$H_0: \gamma_2 = 0$, ingen forskjell i avkastning på utdanning,

$H_A: \gamma_2 < 0$, mindre avkastning på utdanning i Norge.

Vi velger igjen et 5% signifikansnivå og har kritisk verdi på 1,64. t-verdien til *norge_yrs* er -9,10. Vi forkaster H_0 dersom blir absolutt t-verdi større enn kritisk verdi og vi forkaster H_0 . Vi må indirekte akseptere H_A om at avkastningen av utdanning er mindre i Norge.

Ser at γ_1 har økt kraftig fra modell 2 til 3. Dette er grunnet at funksjonen til variabelen *norge* har gått fra å beskrive lønnsnivåforskjeller i landene (modell 2), til å vise forskjellen i konstantleddet mellom landene (modell 3). Det vil si at γ_1 nå vil få en annen funksjon enn i modell2, ved at *norge* nå viser til lønnsforskjeller mellom landene, gitt at individet har null år utdanning. Siden det er tilnærmet null personer som har null år utdanning vil ikke γ_1 være hensiktsmessig å tolke videre i utvidelse av modellen. Observerer også at standardavviket har økt og en høy korrelasjon ($R=0.9721$) mellom γ_1 og γ_2 , som kan være et tegn på multikollinearitet (Thomas, 2005, ss. 462-463). Dette kan gjøre det vanskelig å skille dummiens individuelle effekt på avhengig variabel. Gjennomfører derfor en F-test for å kontrollere at variablene er signifikante og sammen har innvirkning for lønna.

$H_0: \gamma_1 = \gamma_2 = 0$

$H_A: \gamma_1 = \gamma_2 \neq 0$

Med frihetsgrader på 2 i teller og 5155 i nevner, får vi en f-verdi på 188,73 og en p-verdi < 0.01 som vil si at vi må forkaste H_0 til både et 5% og 1% signifikansnivå. γ_1 og avkastning av utdanning har sammen signifikant effekt på timelønn, som vil si at de begge er betydningsfulle for å undersøke forskjellene av avkastning i landene.

Etter denne utvidelsen av modellen ser vi også at forklaringskraften har økt fra 27,66% til 28,79%. Dette er ikke en markant økning og ønsker derfor å undersøke om det er mulig å høyne føyningsgraden ved å tilsette flere kontrollvariabler i kap. 5.5.2.

5.5.2 Alle kontrollvariabler

Ved å legge til flere kontrollvariabler kan reliabiliteten til modellen øke. Vi velger å kontrollere for kjønn, kognitive ferdigheter, yrkesgruppe, sektor og heltid/deltid. I motsetning til dummiene i modell 3 er det ingen interaksjon mellom nasjonalitet og de nye kontrollvariablene. Det betyr at vi ser på innvirkning av X_i , gitt at effekten av X_i er den samme i begge land. Vi ønsker å analysere hva som skjer med den estimerte avkastningsforskjellen, γ_2 , ved en slik utvidelse.

Modell 4:

Modellen utvides først med kognitive ferdigheter.

$$\ln(\text{earnhrppp}) = \beta_0 + \beta_1 \text{yrsqual} + \beta_2 c_q09_c + \beta_3 c_q09_c^2 + \gamma_1 \text{norge} + \gamma_2 \text{norge} * \text{yrsqual} + \beta_5 \text{num1} + \beta_6 \text{lit1} + \beta_7 \text{psl} + \varepsilon_i \quad (5.11)$$

Av Stata kan vi lese at t-verdiene til ferdighetene er høye nok til å ha signifikant individuell innvirkning på lønn ved en tosidig hypotesetest og et signifikansnivå på 0,05. Lese-, skrive- og problemløsningsferdigheter i IKT-miljø, er alle ulike mål på kognitive ferdigheter, og korrelerer sterkt (se korrelasjonsmatrise, tabell V3 i vedlegget). Sterk korrelasjon gjør det vanskelig å separere effekter av ferdighetene og deres individuelle innvirkning på avhengig variabel (Hanushek, Guido, Wiederhold, & Ludger, 2015, s. 108). Ettersom vi har tre variabler i datasettet som måler ferdigheter ønsker vi å undersøke om forskjellige former for evne påvirker lønn og gjennomfører derfor en F-test. Med en nullhypotese om at ingen av ferdighetene har innvirkning på lønn, kan vi forkaste H_0 til en f-verdi på 119,69 med 3 frihetsgrader i teller og 4574 frihetsgrader i nevner. Altså kan vi si at evnene samlet sett har signifikant innvirkning, siden minst én avviker fra antagelsen om null innvirkning.

Vi ser at avkastningen på utdanning reduseres betydelig når vi kontrollerer for kognitive ferdigheter. Avkastningen av utdanning i USA ligger relativt høyt på 9,33 prosent i modell 3 og faller til 7,24 prosent i modell 4. Ser også at det norske avviket fra avkastningen på formell utdanning i USA holder seg relativt stabil. Koeffisienten beveger seg fra å ha en 3,77-

prosentpoeng lavere avkastning av utdanning enn referanselandet, til å få en marginalt mindre avstand på 3,69 prosentpoeng forskjell. Ved t-test til et 5% signifikansnivå vil β_1 og γ_2 fremdeles være signifikante. Altså har ferdighetsvariablene mye å si for avkastningen av formell utdanning i USA, β_1 , men lite å si for avkastningsforskjellene mellom landene, γ_2 . Ved en ny F-test ser vi at effekten av γ_1 og γ_2 fortsatt er signifikante.

Utvidelsen og inkluderingen av flere kontrollvariabler fra modell 3 til modell 4 har justert determinasjonskoeffisient økt fra 28,79% til 31,32%. Vi ønsker likevel å undersøke om forklaringskraften kan øke ytterligere og om virkningen av formell utdanning vil endre seg noe mer.

Modell 5:

Vi utvider modellen ytterligere og inkluderer blant annet mer komplekse indikatorvariabler. Dette er variabler som deles opp i ulike dummyer som representerer forskjellige størrelser eller tallkoder av variabelen. I datasettet vårt benytter vi oss av slike indikatorvariabler for å sjekke for ulike yrkesgrupper (*isco1c*) og i hvilken sektor (*d_q03*) individet tilhører og om disse har noe å si for avkastningen av utdanning mellom landene. For indikatorvariabelen *isco1c* benytter vi yrkeskode 2, som er den best betalte yrkesgruppen, som referansekategori. Det betyr at koeffisienten foran de andre variablene er oppgitt som en prosentandel forskjell relativt til gitt referansekategori. Eksempelvis ser vi at koeffisienten for yrkeskode 10 er -0.4304055. Tolkningen blir da at individer som arbeider i yrker som ikke krever spesifikke kvalifikasjoner tjener tilnærmet 43 prosentpoeng mindre i timen enn ledere fra yrkesgruppe 1. Kontrollerer samtidig for kjønn og fulltid. I likhet med kontrollvariablene fra modell 4 interagerer ikke tilleggsvariablene i modell 5 med landdummy (*norge*) og effekten av faktorene regnes som like i begge land. Får følgende ytterligere utvidet modell:

Modell 5:

$$\ln(\text{earnhrppp}) = \beta_0 + \beta_1 \text{yrsqual} + \beta_2 c_q09_c + \beta_3 c_q09_c^2 + \gamma_1 \text{norge} + \gamma_2 \text{norge} * \text{yrsqual} + \beta_5 \text{num1} + \beta_6 \text{lit1} + \beta_7 \text{psl1} + \text{female} + \text{sektor} + \text{yrke} + \text{fulltime} + \varepsilon_i \quad (5.12)$$

Av regresjon med modell 5 observerer vi at enkelte variable ikke er signifikante. Hvorvidt individet jobber fulltid eller deltid, samt erfaring kvadrert har ikke en signifikant innvirkning på timelønnen til et signifikansnivå på 5%. Allikevel velger vi å beholde disse kontrollvariablene, siden hovedproblemstillingen vår tar utgangspunkt i avkastning av formell utdanning, som gjør at vi heller er interesserte i hvordan γ_2 endrer seg ved tilsettelse av ulike kontrollvariabler.

Leseforståelse (*litscore1*) er heller ikke signifikant i modell 5. Vi kan tolke den lave t-verdien på leseferdigheter som et resultat av høy korrelasjon med de andre ferdighetsvariablene i modellen. Ved å utføre en multipl hypotesetest (F-test) ser vi av Stata at de kognitive ferdighetene fremdeles har en signifikant kollektiv innvirkning på lønn, som ikke nødvendigvis speiles i utdanningen. Dette gjør at vi velger å beholde alle ferdighetsmålene i modellen.

Ser nok en gang at koeffisientene til avkastning av utdanning i USA og Norge har endret seg. Inkludering av variablene kjønn, sektor, yrkesgruppe og heltid fører til en relativt kraftig reduksjon i avkastning av utdanning i begge land, hvor nedgangen er størst i Norge. Selv om vi ser en nedgang i avkastningen av formell utdanning i begge land, vil inkludering av flere kontrollvariabler fortsatt ikke påvirke avkastningsdifferansen mellom landene i betydelig grad. Fra modell 3 til modell 5 observerer vi en vesentlig mindre endring for γ_2 enn for β_1 . γ_2 viser en økning i avkastningsforskjell på 0,39 prosentpoeng, mens β_1 , som viser avkastningen på utdanning i USA, har sunket med 2,45 prosentpoeng. Gjennomfører en ny hypotesetest for å teste om påstandene om β_1 og γ_2 fremdeles er tilstede i modell 5. Både t-test av β_1 og γ_2 , samt f-test av γ_1 og γ_2 viser at de er signifikante. Avkastningen av ett år ekstra høyere utdanning er signifikant større i USA enn i Norge. Dette med en justert forklaringskraft på 40,5% i modellen.

Tabell 5.3: Estimerte resultater i modell (1)-(5). Fullstendige estimeringsresultater i Vedlegg, tabell V2.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	learnhrppp	learnhrppp	learnhrppp	learnhrppp	learnhrppp
yrqual	0.0800*** (0.00213)	0.0760*** (0.00209)	0.0933*** (0.00281)	0.0724*** (0.00342)	0.0575*** (0.00359)
c_q09_c	0.00883*** (0.000511)	0.00898*** (0.000497)	0.00857*** (0.000495)	0.0102*** (0.000528)	0.00841*** (0.000534)
c_q09_c_exp	-5.65e-22*** (2.18e-22)	-4.90e-22** (2.12e-22)	-5.41e-22** (2.11e-22)	-3.59e-22 (2.52e-22)	-5.13e-22 (2.69e-22)
norge		0.188*** (0.0110)	0.739*** (0.0616)	0.746*** (0.0683)	0.817*** (0.0674)
norge_yrs			-0.0377*** (0.00414)	-0.0369*** (0.00452)	-0.0416*** (0.00447)
Kontrollert for ferdigheter	Nei	Nei	Nei	Ja	Ja
Kontrollert for kjønn, fulltid, bransje og sektor	Nei	Nei	Nei	Nei	Ja

Observations	5,161	5,161	5,161	4,583	4,226
Adjusted R-squared	0.236	0.277	0.288	0.313	0.405

Standardavvik i parentes

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

5.6 Heteroskedastisitet

Som vist i teorikapittelet, setter OLS-forutsetningene (i)-(iv) sterke restriksjoner på modellen. For å teste kvaliteten på modellen velger vi å teste for heteroskedastisitet. Det er flere metoder for å teste for heteroskedastisitet, der en Lagrange- multiplikator eller en Breusch-Pagan test er de mest brukte (Thomas, 2005, s. 505). Testen undersøker at forutsetning (ii) om konstant varians i restleddet for alle i holder: $Var(\varepsilon_i) = \sigma^2$. En av konsekvensene av heteroskedastiske støyledd er at formelen for estimatorens standardavvik blir feil (Thomas, 2005, s. 480). For å kunne trekke konklusjoner om effekten av hva eksempelvis ett år ekstra utdanning vil ha å si for lønna, må det være konstant varians på feilleddene. I tillegg er lik varians avgjørende for få korrekte konfidensintervall og å utføre hypotesetester. Velger å anvende Breusch-Pagan-testen for å teste modellen vår for heteroskedastisitet, grunnet at denne kan benyttes i Stata. Testmetoden går ut på å undersøke om variansen i støyleddene fra en lineær regresjon er betinget av verdiene i de uavhengige variablene. W_i betegner variabler som vi tror kan ha innvirkning på variansen. a_i er konstanter.

$$Var(\varepsilon_i) = f(a_1 + a_2W_2 + \dots + a_mW_m) \quad (5.13)$$

Velger å teste modell 5 for heteroskedastisitet, for å kunne undersøke om slutningene vi har trukket stemmer eller ikke:

$H_0: a_2 = a_3 = \dots = a_m = 0$, hvor restleddet er homoskedastisk

$H_A: a_2 = a_3 = \dots = a_m \neq 0$, hvor restleddet er heteroskedastisk

Breusch-Pagan-testen gir følgende testobservator med χ^2 -kvadratfordeling med $(m-1)$ frihetsgrader, hvor m =antall konstanter (a): (Thomas, 2005, ss. 482-483).

$$TS = nR^2 \sim \chi^2(m-1) \quad (5.14)$$

Breusch-Pagan testen finner predikerte verdier (\hat{Y}) og residualene (e). Deretter blir residualene kvadrert og omskalert slik at gjennomsnittet er 1. De kvadrerte residualene blir deretter regressert for \hat{Y} . Dersom nullhypotesen er sann finnes det ingen heteroskedastisitet og testen har en χ^2 -kvadratfordeling med én frihetsgrad. (Williams, 2015, s. 9)

Tester en *kji*-kvadratfordelt nullhypotese om at variansen er konstant og at vi har homoskedastiske støyledd. Med en *kji*-kvadratverdi lik 11,96 får vi en *p*-verdi på 0,0005 som gjør at vi forkaster nullhypotesen til et 5% signifikansnivå og konkluderer med at variansen er ulik. Vi tester altså positivt for heteroskedastisitet, noe som i utgangspunktet ugyldiggjør resultatene funnet ved hjelp av OLS. Dette er et brudd på forutsetning (ii) om konstant varians i restleddet, men ved å estimere robuste standardfeil ser vi at endringene er ingen eller svært små. Derfor slutter vi at hypotesetestene fortsatt er signifikante (Thomas, 2005, ss. 498-505):

Variabel	T-verdi	T-verdi r.	Endring	Koeffisient	Koeffisient r.	Endring
yrsqual	16.02	13.25	2.77	0.0575485	0.0575485	0
c_q09_c	15.76	14.52	1.24	0.0084146	0.0084146	0
c_q09_c^2	-1.62	-1.23	0.39	-4.29e-22	-4.29e-22	0
norge	12.12	12.26	0.14	0.8172186	0.8172186	0
norge_yrs	-9.32	-9.16	0.16	-0.0416273	-0.0416273	0
litscore1	-0.45	-0.44	0.01	-0.0061543	-0.0061543	0
numscore1	4.08	3.88	0.2	0.0540644	0.0540644	0
pslscore1	3.44	3.42	0.02	0.0343988	0.0343988	0
female	-8.39	-8.05	0.34	-0.1012943	-0.1012943	0
full_time	1.72	1.45	0.27	0.0293309	0.0293309	0

Ser at det er ingen endring i koeffisientene, og relativt små endringer i *t*-verdiene. Heteroskedastisiteten skal være betydelig stor for å gjøre alvorlig utslag på standardfeilene. De små endringene stemmer også for variablene *isco1c* og *d_q0*. Vi ser på grunnlag av dette bort i fra ulik varians og små endringer, og vi beholder modellen.

5.7 Tolkning av resultat

I den deskriptive fremstillingen av vårt datamateriale ser vi at gjennomsnittslønnen til gitt utdanningsnivå og yrkesgruppe (tabell 4.4 og 4.5) har en større spredning i USA enn i Norge. Dette kan tyde på at Norge har større lønnsutjevning og at bransjetilhørighet kan ha mindre påvirkning på lønna i Norge enn i USA, noe som kan ha med grad av fagorganisering å gjøre. Disse tendensene får vi derimot ikke testet i vår regresjonsanalyse, fordi vår modell er for restriktiv ved at den ikke skiller mellom Norge og USA i disse variablene.

Tallene i regresjonsanalysen viser at inkludering av kognitive ferdigheter, fulltid, kjønn, bransje- og sektorforskjeller nesten ikke endrer avkastningsforskjellene mellom Norge og USA, kun avkastningsnivået av formell utdanning i landene. Det er et interessant funn at forskjellen er forholdsvis stabil og holder seg robust gjennom utvidelse av modellen. Dette kan tyde på at

det er de strukturelle forskjellene mellom USA og Norge som ligger til grunn for avkastningsforskjellene av formell utdanning γ_2 . De strukturelle forskjellene kan, som nevnt innledningsvis, komme av ulik grad av fagorganisering.

Påstanden om at forskjellen mellom avkastning av formell utdanning kun skyldes at Norge har sterke fagforeninger og sammenpresset lønnsstruktur kan ikke påvises med disse funnene. Derimot kan funnene tyde på at sentraliserte lønnsforhandlinger kan være en årsak til avkastningsforskjellene. En annen årsak til avkastningsforskjellene kan for eksempel være svært forskjellige utdanningssystem, som bidrar til andre kostnader og sosiale faktorer som ikke inkluderes i modellen (Patrinos, 2016). USA baserer seg mye på privatfinansiering av formell utdanning, mens Norge har et langt mer offentlig finansiert skolesystem. Som empirien sier forutsetter privatfinansiering av skolesystemet at avkastningen av formell utdanning må være høy for at individer skal ha incentiver til å ta høyere utdanning i USA. Sannsynligvis finnes det også andre strukturelle mekanismer som har innvirkning på avkastningsforskjellene. I modellene vil disse kun fanges opp i restleddet som ved bruk av OLS vil estimeres til null. For å kontrollere virkningen av slike faktorer trenger man et mer omfattende datamateriale som vi i denne oppgaven ikke har tilgang på.

6. Oppsummering og konklusjon

6.1 Oppsummering

I denne oppgaven har vi forsøkt å undersøke om det foreligger avkastningsforskjeller av formell utdanning i Norge og USA med utgangspunkt i datamaterialet fra PIAAC-undersøkelsen fra 2012. Vi har sett på hvordan sektor, kjønn, arbeidserfaring, kognitive ferdigheter, og stillingsprosent har innvirkning på økonomisk avkastning. For å undersøke dette har vi først benyttet tidligere empiri, blant annet Jacob Mincers inntektsfunksjon og Gary Beckers humankapitalteori, samt teori om hvordan fagorganiseringsgrad har påvirkning på lønn.

Vi har brukt OLS (minste kvadraters metode) som metode, hvor vi anvendte en log-lineær inntektsfunksjon som utgangspunkt for vår analyse for å se på prosentvis endring i koeffisientene og avhengig variabel.

Først brukte vi en restriktiv modell med utgangspunkt i Mincers inntektsfunksjon der *learnhrppp* var den avhengige variabelen, Y . Modellen kontrollerte for hvordan formelt utdanningsnivå (*yrsqual*), erfaring (*c_q09_c*) og en ikke-lineær effekt av erfaring (*c_q09_c_2*)

påvirket det gjennomsnittlige lønnsnivået for landene samlet sett. Deretter utvidet vi modellen til å kontrollere for land, samt et interaksjonsledd for å kunne se på avkastningsforskjeller. Til slutt anvendte vi kontrollvariabler for ferdigheter, kjønn, fulltid, yrkesgruppe og sektor for å undersøke om disse hadde noen effekt på avkastningsforskjellene.

Avslutningsvis i analysen måtte vi teste for heteroskedastisitet og homoskedastisitet, altså hvorvidt variansen i datasettene samsvarer (homo) eller ikke (hetero). Vi måtte forkaste nullhypotesen om homoskedastisitet som gjorde at vi måtte korrigere for heteroskedastisitet. Slik korrigerer gjøres ved å estimere robuste standardfeil. Observerte at justeringene i koeffisientene og t- verdiene var små og vi valgte derfor å beholde vår modell.

6.2 Konklusjon

Oppgavens hovedfunn er at det finnes signifikante forskjeller i avkastning for formell utdanning mellom Norge og USA. I vår fullstendige modell kontrollert for alle valgte kontrollvariabler finner vi at avkastningen for ett års ekstra utdanning i USA er på 5.8%, mens et individ i Norge har en avkastning på 1.59%. Dette avkastningsgapet holder seg relativt stabilt gjennom hele vår analyse. Som sagt kan dette tyde på at avkastningsforskjellene skyldes strukturelle forskjeller mellom landene, noe som underbygger tidligere forskning som peker på at sentraliserte lønnsforhandlinger og sterk fagorganisering kan føre til lavere avkastning på utdanning.

7. Referanser

Bjørkeng, B. (2013). *Ferdigheter i voksenalderen*. Oslo - Kongsvinger: Statistisk Sentralbyrå.

FN-Sambandet. (2017). *BNP per innbygger*. Hentet fra FN-Sambandet :
<https://www.fn.no/Statistikk/BNP-per-innbygger>

Hanushek, E. A., Guido, S., Wiederhold, S., & Ludger, W. (2015, January). Returns to skills around the world: Evidence from PIAAC. *European Economic Review* , ss. 103-130.

Hægeland, T. (2003). *Økonomisk avkastning av utdanning*. Hentet fra Statistisk Sentralbyrå :
<https://www.ssb.no/a/publikasjoner/pdf/sa60/kap-10.pdf>

Mincer, J. (1958, August). Investment in Human Capital and Personal Income Distribution. *Journal of Political Economy* , ss. 281-302.

OECD. (2018). *Education at a Glance 2018: OECD Indicators*. Paris: OECD Publishing. Hentet fra
<https://doi.org/10.1787/eag-2018-en>

Patrinos, H. A. (2016). *Estimating the return to schooling using the Mincer equation*. Washington, D.C.: IZA World of Labor.

Thomas, R. L. (2005). *Using Statistics in Economics*. Berkshire: McGraw-Hill Education.

Williams, R. (2015). *Heteroskedasticity*. Paris: University of Notre Dame.

Vedlegg

Tabell V1: Bransjefordeling, *isco1c*-variabelen

Verdi	Bransje/yrkesgruppe
2	Lovgivende yrker, høytstående tjenestemenn og ledere
3	Fagfolk
4	Teknikere og partnere
5	Kontorarbeider
6	Servicemedarbeidere og butikkansatte
7	Faglærte innen landbruk og fiske
8	Håndverksarbeidere
9	Anleggs- og maskinoperatører
10	Elementære yrker
11	Ikke lønnet arbeid de siste 5 årene

Tabell V2: Fullstendig oversikt over estimerte resultater i modell (1)-(5).

VARIABLES	(1)	(2)	(3)	(4)	(5)
	learnhrppp	learnhrppp	learnhrppp	learnhrppp	learnhrppp
yrsqual	0.0800*** (0.00213)	0.0760*** (0.00209)	0.0933*** (0.00281)	0.0724*** (0.00342)	0.0575*** (0.00359)
c_q09_c	0.00883*** (0.000511)	0.00898*** (0.000497)	0.00857*** (0.000495)	0.0102*** (0.000528)	0.00841*** (0.000534)
c_q09_c^2	-5.65e-22*** (2.18e-22)	-4.90e-22** (2.12e-22)	-5.41e-22** (2.11e-22)	-3.59e-22 (2.52e-22)	-4.29e-22 (2.64e-22)
norge		0.188*** (0.0110)	0.739*** (0.0616)	0.746*** (0.0683)	0.817*** (0.0674)
norge_yrs			-0.0377*** (0.00414)	-0.0369*** (0.00452)	-0.0416*** (0.00447)
litscore1				-0.0380*** (0.0137)	-0.00615 (0.0138)
numscore1				0.105*** (0.0129)	0.0541*** (0.0133)
pslscore1				0.0612*** (0.0100)	0.0344*** (0.0100)
female					-0.101*** (0.0121)
full_time					0.0293* (0.0170)
3.isco1c					-0.0670*** (0.0199)
4.isco1c					-0.108*** (0.0206)

5.isco1c					-0.317*** (0.0268)
6.isco1c					-0.371*** (0.0230)
7.isco1c					-0.273** (0.112)
8.isco1c					-0.208*** (0.0274)
9.isco1c					-0.299*** (0.0335)
10.isco1c					-0.430*** (0.0371)
2.d_q03					-0.0678*** (0.0123)
3.d_q03					-0.0953*** (0.0256)
Constant	1.728*** (0.0346)	1.674*** (0.0338)	1.435*** (0.0426)	1.691*** (0.0512)	2.184*** (0.0612)
Observations	5,161	5,161	5,161	4,583	4,226
Adjusted R-squared	0.236	0.277	0.288	0.313	0.405

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Tabell V3: Korrelasjonsmatrise

	learnh rppp	yrsqua l	C_q09 _c	C_q09 _c^2	norge	Norge _yrs	litscor el	numsc orel	pslsco rel	femal e	Full_ti me	isco1c	D_q0 3
learnh rppp	1.000 0												
yrsqua l	0.399 6	1.000 0											
C_q09 _c	0.189 5	- 0.088	1.000 0										
C_q09 _c^2	0.026 5	- 0.016	0.258 1	1.000 0									
norge	0.219 5	0.068 2	- 0.046	- 0.055	1.000 0								
Norge _yrs	0.252 0	0.216 1	- 0.073	- 0.063	0.976 0	1.000 0							
litscor el	0.295 2	0.444 8	- 0.169	- 0.044	- 0.050	0.014 7	1.000 0						
numsc orel	0.352 9	0.445 0	- 0.094	- 0.026	- 0.044	0.019 3	0.860 6	1.000 0					
pslsco rel	0.290 1	0.402 4	- 0.227	- 0.048	- 0.037	0.025 2	0.822 9	0.785 0	1.000 0				
femal e	- 0.194	0.047 4	- 0.070	- 0.055	- 0.052	- 0.041	- 0.036	- 0.165	- 0.070	1.000 0			
Full_ti me	0.133 1	0.056 9	0.026 4	- 0.003	- 0.050	- 0.033	0.097 6	0.121 5	0.105 6	- 0.203	1.000 0		

isco1c	-	-	-	-	-	-	-	-	-	-	-	1.000	
	0.387	0.505	0.039	0.006	0.000	0.077	0.357	0.340	0.352	0.118	0.104	0	
	0	3	9	5	9	1	8	1	4	6	2		
D_q0	-	0.238	0.021	0.001	-	0.018	0.074	0.026	0.026	0.193	-	-	1.000
3	0.001	0	4	3	0.014	4	9	7	6	7	0.091	0.218	0
	8				0						6	5	

