Bacheloroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
Institutt for samfunnsøkonomi

Lars Djuve Båtsvik
Ingrid Bergly Pettersen
Sigve Naustdal Schjølset
Amanda Bakken Sune

Leseferdigheter til barn med utenlandske foreldre

En kvantitativ analyse

Bacheloroppgave i Samfunnsøkonomi
Veileder: Bjarne Strøm

Mai 2020

NTNU
Kunnskap for en bedre verden

Innholdsfortegnelse

1. Innledning

1.1 Aktualisering av problemstillingen

I 2018 avga Stoltenbergutvalget en rapport som tok for seg kjønnsforskjeller i skoleprestasjoner og utdanningsløp. Rapporten viser at gutter gjør det dårligere enn jenter, dette gjelder spesielt tidlig i studieløpet (Borgonovi, Ferrara, & Maghnouj, 2018, s. 106-109). At gutter presterer dårligere enn jenter på skolen har lenge vært kjent. Dette skaper ulike forutsetninger for videre utdanning og senere karriereliv. Spesielt store forskjeller ser vi i regne- og leseferdigheter. Gjennom rapporten legger ekspertutvalget frem tiltak som skal gi barn like muligheter i skolen helt uavhengig av kjønn og sosial bakgrunn. For vår problemstilling er denne rapporten spesielt interessant grunnet funnene av hvilke faktorer som spiller inn for forskjellen i elevprestasjonene.

Tiltakene som blir anbefalt i rapporten har som hensikt å redusere de sosioøkonomiske forskjellene som viser seg å gi utslag for elevprestasjonene. I likhet med hva vi skal drøfte i vår artikkel blir det lagt vekt på hvilken påvirkning familie og oppvekstmiljø har for elevprestasjonene. Viktigheten av tilpasset undervisningsform som skal skape samme forutsetninger for barn uavhengig av sosial bakgrunn kan ikke understrekes nok. Samfunnet er pliktet gjennom barneloven og FNs barnekonvensjon til å tilpasse undervisningen og oppvekstsvilkårene til alle barn så de får rettferdige forutsetninger for videre liv (familiedepartementet, 2000).

Studier av frafall etter grunnskolen (1-10.trinn) viser at det får store samfunnsøkonomiske konsekvenser når flere velger å ikke ta videre opplæring. Det er vist at individer som har fullført videre opplæring er mer produktive og har en positiv påvirkning på kollegers produktivitet på arbeidsplassen. De positive ringvirkningene av høyere utdanning sparer samfunnet for store samfunnsøkonomiske laster. En rapport fra 2010 viser hvordan en økning fra 70% til 80% av elever som fullfører videregåendeskole kan gi samfunnet en gevinst på 5,4 milliarder per kull (Falch, Johannesen, & Strøm, 2009, s. 49-50).

Fullført skolegang gir individer en avkastning i form av lønn. Kostnaden er tapt arbeidsgevinst for årene brukt på utdanning. Nyttegevinsten av skolegang vil være klart høyere enn tapet av arbeidsgevinst. Risikoen for å falle utenfor i rus- eller kriminelle miljøer reduseres betraktelig ved økt utdanning. Viktigheten av å mestre grunnleggende ferdigheter på skolen for å motiveres til videre utdanning er viktig for individers egen fremtid og for samfunnet som helhet (Camilla Stoltenberg, 2019, s. 217-232; Falch et al., 2009, s. 49).

1.2 Presentasjon av problemstilling

Denne artikkelen skal undersøke leseresultater blant norske barn. Mer spesifikt ønsker vi å se på hvordan barn av utenlandske foreldre gjør det i forhold til elever med foreldre som er født i Norge. Vi har utformet problemstillingen:

"Er det en negativ sammenheng mellom leseferdigheter og det å ha utenlandske foreldre?"

Det korte svaret er "ja". Det er en negativ sammenheng mellom testresultater i lesing og det å ha utenlandske foreldre. Dette kan bunne i en rekke sosioøkonomiske årsaker. Vi ønsker å se i hvilken grad disse årsakene påvirker leseferdigheter. Vi vil også se nærmere på om vi kan trekke en årsakssammenheng mellom leseferdigheter og det å ha utenlandske foreldre. Vi har kommet fram til en nullhypotese og en alternativhypotese som vi ønsker å teste. Gjennom hypotesetesten vil vi undersøke om det er grunnlag for å forkaste eller beholde nullhypotesen.

H_0: Det e**r ikke** en negativ korrelasjon mellom *leseresult* og *utlandForld*

H_A: Det **er** en negativ korrelasjon mellom *leseresult* og *utlandForldr*

For å teste denne har vi satt opp tre modeller som i ulik grad kontrollerer for sosioøkonomiske faktorer. Våre funn sier at vi ved bruk av en restriktiv modell kan forkaste H0, men dette blir feil da bruken av en mer fullstendig modell forteller oss at vi ikke har statistisk grunnlag til å forkaste H0.

Det er bred konsensus om at jenter presterer bedre på skolen enn det gutter gjør (Borgonovi et al., 2018). I tillegg er det gjort studier som sier at gutter er mer ømfintlige for miljøendringer (Ibid). Derfor tenker vi at det er interessant å utforske om effekten av ikke-norske foreldre er større for gutter enn for jenter som en utvidet problemstilling.

1.3 Metode

For å svare på problemstillingen vil vi utføre en empirisk analyse der vi tar i bruk minste kvadraters metode (MKM) med data hentet fra PIRLS 2001. Det hadde vært mulig å gjøre en kvalitativ undersøkelse ved å gjøre intervjuer eller observasjoner av et lite utvalg skoleklasser, men dette ville tatt lang tid og krevd mye arbeid. Heldigvis er PIRLS-datasettene gode datasett med et stort utvalg. Det vil derfor være mer hensiktsmessig å ta i bruk kvantitativ forskningsmetode for å analysere problemstillingen vår.

MKM gir oss mulighet til å undersøke sammenhengen mellom en avhengig variabel og flere forklaringsvariabler/responsvariabler. Estimeringsmetoden egner seg godt for å analysere problemstillingen vår. Grunnen til dette er at den gir oss muligheten til å analysere sammenhengen mellom leseferdigheter og effekten av utenlandske foreldre mens vi kontrollerer for flere andre uavhengige variabler. Siden vi gjør en kvantitativ analyse, vil dataene vi bruker bestemme utfallet av undersøkelsen. Det er derfor svært viktig at man velger gode data. Mengden data vil være avgjørende for å oppnå presise resultater. Datasettet fra PIRLS-undersøkelsen fra 2001 har over 3000 respondenter, noe som er mer enn tilstrekkelig (Kotrlik & Higgins, 2001, s. 48).

2. Teoretisk rammeverk og tidligere litteratur

For vår analyse vil vi benytte oss av en skoleproduktfunksjon. Skoleproduktfunksjoner blir ofte brukt i analyser og forskningsprosjekter som undersøker hvilke faktorer som bidrar til bedre skoler. Det som gjør funksjonen til en skoleproduktfunksjon er at vi ser på skoletestresultater. Tidligere var det vanlig å se på lønninger senere i arbeidslivet eller innsatsfaktorer i skolen, som for eksempel hvor mye penger det ble brukt per elev (Hanushek, 2020).

Vedrørende elevprestasjonene i Norge blir det gjort kontinuerlige målinger for å kartlegge og sikre kvaliteten i grunnskolen. Nasjonale prøver brukes som redskap til å teste elevers lese- og regneferdigheter. Historisk har det skjedd en vending i hvordan skoleferdighetene måles. Kvaliteten på skolen har gått fra å bli målt i ressurser som blir satt inn i skolen, til hvilke ressurser som kommer ut. Det vil si elevers sluttresultater. Vi skal i denne artikkelen finne ut om barn av utenlandske foreldre gjør det dårligere i lesing sammenlignet med barn av norske foreldre.

2.1 Teoretisk rammeverk

Vi vil bruke en skoleproduktfunksjon i vår analyse. Denne produktfunksjonen stammer fra en mer generell analyse om nivået på humankapital. Tradisjonelt har forskning på humankapital fokusert på implikasjoner for arbeidsmarkedet og lønnsfastsettelse. I de senere årene har man med skoleproduktfunksjonen lagt vekt på faktorene som bestemmer ferdighetsnivået til individene som utgjør humankapitalen (Hanushek, 2020, s. 161-162).

I 1966 kom Colemanrapporten. Sosiologen James Colemans oppdrag var å undersøke empirisk om skolen kunne brukes til å redusere raseforskjellene i USA ved å styrke

utdannelsen for den svarte befolkningen (Coleman, 1968). Det ble konkludert i rapporten at skolen ikke kunne påvirke utfallene (Bonesrønning, 2004, s. 16). Forskerteamet bak rapporten pekte på at den viktigste indikatoren for skoleprestasjoner var verken skolebygg eller ressurser, men familie (Dickinson, 2016). Rapporten rettet fokus mot elevenes prestasjoner, output, i motsetning til utgifter per elev og lærerkarakteristika, input. Vi kan her trekke et skille mellom to typer input som bestemmer elevenes output: variabler som er kontrollerbare for myndighetene (karakteristika ved skolene, lærere og pensum) og variabler som er utenfor myndighetenes kontroll (familiekarakteristika, peer group-effects) (Hanushek, 2020, s. 164).

Motivasjonen bak forskningen på denne typen faktorer handler om hvordan ferdighetene til humankapitalen kan utvikles og forbedres. På denne måten har man vist at det kan få potensielt store (positive) økonomiske konsekvenser å utvikle analyser om skoleprestasjoner. Konsekvensene gjør seg synlige i form av produktivitetsnivået i arbeidsmarkedet, her ser vi altså en sammenheng mellom skoleprestasjoner og arbeidsmarkedet. (Hanushek, 2020, s. 161-162).

2.2 Litteratur og tidligere studier

Arne Lervåg og Monica Melby-Lervåg har gjennomført en metaanalyse med tittelen «Muntlig språk, ordavkoding og leseforståelse hos tospråklige: En sammenfatning av empiriske studier» fra 2009. En metaanalyse er bruk av statistiske metoder for å sammenligne resultater fra studier med samme problemstilling (Frøslie 2017). Her har de inkludert data og resultater fra 34 studier, med mange uavhengige effektstørrelser som måler muntlig språk, ordavkoding og leseforståelse. (Lervåg & Melby-Lervåg 2009) Det blir avdekket at muntlige språkferdigheter hos de tospråklige elevene er klart svakere enn hos de enspråklige elevene (Melby-Lervåg & Lervåg, 2009, s. 12).

Når det gjelder leseferdigheter trekker forskerne et skille mellom ordavkoding og leseforståelse. *Ordavkoding brukes om ulike strategier som gjør leseren i stand til å identifisere ordets uttale og mening, og forutsetter at leseren har knekket den alfabetiske koden* (Gabrielsen & Lundetræ 2013, s. 119). De tospråklige elevene viser seg å ha like gode ordavkodingsferdigheter som de enspråklige elevene. I tillegg scorer tospråklige faktisk bedre på ordavkoding sammenlignet med enspråklige i tidlig alder (Melby-Lervåg & Lervåg 2009, s. 13-14). På grunnlag av dette kan forskerne konkludere med at de svake leseferdighetene henger sammen med svake språkferdigheter.

2.2.1 Liv Bøyesen

Liv Bøyesen understreker også viktigheten av elevens språklige ferdigheter når de leser. Hun gjengir Lise Iversen Kulbrandstads uttrykk "å lese med øret". Dette er en fin metafor for å beskrive hvordan flerspråklige barn har dårligere leseferdigheter. (Bøyesen, 2004) Minoritetsbarna ligger bak enspråklige barn når det kommer til de språklige erfaringene. Tospråklige barn kan blant annet ha ulike assosiasjoner til omverdenen enn enspråklige barn. Et eksempel kan være at enspråklige norske barn vet at solen heter sol, og dermed assosierer ordet og tingen med bokstaven "s". Men for et barn med for eksempel somalisk som morsmål så heter sol "qorrax". Det gjør at barnet assosierer tingen sol med bokstaven og lyden "q". Det vil derfor ta lengre tid før et tospråklig barn vil koble ordet sol med tingen sol. Dette er en av utfordringene som tospråklige barn møter på. Barns språklige erfaringer gjør et stort utslag for leseferdighetene. Om elevene ikke eksponeres for norsk i hjemmet vil det kunne ha negative utslag (ibid.).

2.2.2 Nasjonale prøver

Alice Steinkellner fra Statistisk Sentralbyrå bekrefter funnene ovenfor i sin artikkel "Hvordan går det med innvandrere og deres barn i skolen?" fra 2016. Ved bruk av nasjonale prøver har hun gjort funn om at både innvandrere og norskfødte med innvandrerforeldre presterer dårligere enn de øvrige elevene på de fleste prøver (Steinkellner, 2017). Dette kan skyldes de samme faktorene som er nevnt ovenfor, nemlig hvilket språk som snakkes hjemme.

I en omfattende artikkel har Minja Tea Dzamarija gjort en rekke interessante funn (Dzamarija, 2016, s. 63-64). Gjennom undersøkelser av inntekt, innvandrerbakgrunn, demografi, kjønn og utdanning fant hun i gruppen av elever med innvandrerbakgrunn at: I: jenter gjør det bedre enn gutter i alle aldersgrupper. II: norskfødte med innvandrerforeldre er mye mer sannsynlig til å gjennomføre VGS enn innvandrere. Dette skyldes i stor grad språkbruk i hjemmet (ibid.).

2.4 Oppsummering

Studiene nevnt ovenfor bruker alle en skoleproduktfunksjon for å analysere skoleresultater eller skolens produkt. På lik linje med Coleman analyseres elevprestasjoner ved å knytte testresultater opp mot ulike forhold som går på medelever eller familie. Som vi ser spiller skoleproduktfunksjonen en sentral rolle når man studerer effekten av skole, og det er bred konsensus hos forskere om at sosioøkonomiske forhold spiller en viktig rolle for elevprestasjoner. Innvandrere, gutter, familier med lav inntekt og familier med lav utdanning

er de som kommer dårligst ut (Borgonovi et al., 2018; Bøyesen, 2004; Coleman, 1968; Melby-Lervåg & Lervåg, 2009).

3. Empirisk strategi

I dette kapittelet skal vi presentere, og gå gjennom minste kvadraters metode. Videre forklares det nærmere hva korrelasjon sier oss, og hva vi bruker hypotesetesting til.

3.1 Økonometrisk teori

3.1.1 MKM

Vi vil benytte MKM, minste kvadraters metode, for å vise en eventuell årsakssammenheng mellom barn med utenlandske foreldre og deres leseresultater. Ifølge Thomas er MKM den beste metoden vi vet for å estimere en linje til et spredningsplott. (Thomas, 2005 s. 266). For å utføre en regresjonsanalyse forutsettes antakelsen om lineær årsakssammenheng mellom variabler. I metoden estimeres ukjente parametere i en lineærregresjonsmodell, altså den avhengige utfallsvariabelen, Y, og en eller flere uavhengige variabler, X_i. Det er ikke gitt at forventningsverdien til Y, E(Y), er lik den faktiske verdien på Y. Avviket mellom E(Y) og faktisk Y er gitt ved støyleddet, ε. Et slikt avvik vil forekomme på grunn av uforklarte, og ofte umålbare faktorer som spiller inn på forventningsverdien til populasjonen.

MKM krever fravær av eksakt sammenheng mellom x-ene, og forlanger samtidig sterke forutsetninger om restleddet. For at resultatene ikke skal være misvisende, må betingelsene nedenfor være oppfylt (Studenmund, 2017, s. 111; Thomas 2005, s. 356-359).

I. Regresjonsmodellen er lineær, har riktig funksjonsform og har et støyledd
II. Støyleddet har 0 i populasjonsgjennomsnitt
III. Alle uavhengige variablene er ukorrelert med støyleddet
IV. Observasjon av støyleddet er ukorrelert med hverandre
V. ε har en konstant varians (ingen heteroskedastisitet)
VI. Ingen av de uavhengige variablene er en perfekt lineær funksjon av en annen uavhengig variabel (ingen perfekt multikollinearitet)
VII. ε_i er normalfordelt. (dette er viktig når vi skal teste hypoteser)

Ved bruk av MKM minimeres summen av variansen, som er kvadratet til avviket mellom den estimerte og den observerte verdien. Dette gir oss estimater på både konstantleddet, α, og stigningstallet, β. Dermed får vi en estimert regresjonslinje:

$Yhatt=a + bX_i$ (3.2)

3.1.2 Korrelasjon

For å se på grad av samvariasjon mellom to eller flere variabler beregner vi korrelasjonskoeffisienten, ro. Ved å beregne korrelasjonen i et utvalg får vi et estimat på hvordan korrelasjonskoeffisienten ser ut. Utvalgets korrelasjonskoeffisient er gitt ved:

$$R = \frac{\sum_{i=0}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n} (x_i - \bar{x})^2 \sum_{i=0}^{n} (y_i - \bar{y})^2}}$$

Når R = –1, er korrelasjonen perfekt negativ. Hvis den ene variabelen øker med én prosent depresierer den andre variabelen med en prosent.

Når R = 0, er det ingen korrelasjon mellom variablene. En endring i den ene variabelen påvirker ikke den andre variabelen.

Når R=1, er variablene perfekt positivt korrelert. En økning på én prosent i den ene variabelen fører til en prosent økning i den andre variabelen.

En perfekt sammenheng mellom X-ene stemmer ikke overens med MKM-forutsetningene. Det blir umulig å estimere separate effekter av ulike X-er og å finne separate β-er, kun en estimert kombinasjon. (Thomas, 2005, s. 257-258).

Det er viktig å være klar over at korrelasjon er et mål på samvariasjon, og ikke nødvendigvis sammenheng. Selv om to variabler korrelerer med hverandre er det ikke nødvendigvis slik at sammenhengen er kausal. Spuriøse årsakssammenhenger kan oppstå i en situasjon der en tredje variabel, Z, påvirker sammenhengen mellom variablene (Thomas, 2005, s. 258).

Estimatene gir oss den beste mulige linjen, men vi ønsker å beregne hvor godt regresjonslinjen faktisk beskriver datasettet. Det gjør vi ved å beregne føyningsmålet som uttrykkes ved determinasjonskoeffisienten, R^2. Denne koeffisienten gir oss et mål på hvor stor andel av variasjonen i Y, SST, som kan forklares av variasjonen i X, SSE. Med andre ord, andelen forklart variasjon (SSE) i forhold til total variasjon (SST) (Thomas, 2005, s.

274-276). Justert R^2 måler prosentandelen av variasjon i Y rundt gjennomsnittet som er forklart av regresjonsmodellen, justert for frihetsgrader (Thomas, 2005, s. 277).

$$R^2 = \frac{SSE}{SST} = \frac{b^2 \sum x_i^2}{\sum y_i^2}$$

Når $R^2 = 0$ har modellen ingen forklaringskraft

Når $R^2 = 1$ vil modellen kunne forklare datasettet perfekt.

Siden R^2 har en verdi på mellom 0 og 1 vil forklaringskraften nesten alltid øke desto flere varianter av X som inkluderes i modellen. Dette vil ikke alltid stemme og det gjør at R^2 kan gi et misvisende mål. Derfor benytter vi heller en justert versjon av determinasjonskoeffisienten, som tar hensyn til at flere variabler tilføyes modellen, og justerer forklaringskraften deretter. (Thomas, 2005, s. 421).

3.1.3 Hypotesetesting

Hypotesetesting anvendes med hensikt om å avgjøre hvorvidt MKM-estimatorene stemmer overens med virkeligheten ved en gitt prosentandel usikkerhet. Usikkerheten betegnes som signifikansnivået. Ved utførelse av hypotesetesten kreves en nullhypotese (H_0) og en alternativhypotese (H_A). H_0 er den hypotesen man vil undersøke hvorvidt det er grunnlag for å forkaste, mens H_A vil være en komplementær hypotese man ønsker å underbygge (Thomas, 2005, s. 369-371). Testobservatoren (TS), sammen med kritisk verdi, avgjør hvilke verdier som kan forkaste H_0. Det brukes ofte et 5% signifikansnivå på statistiske tester. Det betyr at det er 5% sjanse for å gjøre en type 1 feil, å forkaste en sann nullhypotese. Motsatt betyr type 2 feil at man beholder en usann nullhypotese. Sannsynligheten for at nullhypotesen er sann vises i p-verdien.

Testene vi skal ta i bruk er t-testen og F-testen. En t-test benyttes for å undersøke om effekten av enkeltparametere har på den avhengige variabelen er signifikant når vi har ukjent populasjonsvarians (Thomas, 2005, s. 428).

F-testen brukes til å teste multiple hypoteser, for å sjekke om de har den samme kollektive signifikante effekten på avhengig variabel. Her er vi ute etter endringen i residualvariasjonen (SSR) når restriksjoner blir lagt til i en modell (Thomas, 2005, s.439-443). En stor økning i SSR indikerer at restriksjonen er ugyldig og at H_0 må forkastes.

3.2 Oppsummering

Kapittelet gir en innføring i MKM, og relevante koeffisienter, R og R^2. Siste delkapittel handler om hvordan en hypotesetest utføres, og kort om hva vi tester ved t- og F-test.

4. Datamaterialet, presentasjon av data

Vi bruker PIRLS (2001) som består av data om 4.klassingers leseferdigheter. I dette kapittelet vil vi presentere variablene som legges til grunn for analysen og forklare hvorfor vi har valgt å omkode fire kategorivariabler til dummyvariabler. Videre følger deskriptiv statistikk for avhengig variabel, interessevariabel og kontrollvariabler samt en tabell som viser spredningen til vår eneste kategorivariabel inntekt. I tillegg presenterer vi en enkel korrelasjonsmatrise som viser sammenhengen mellom samtlige variabler i modellen. Til slutt drøftes det rundt styrker og eventuelle svakheter knyttet til datasettet.

4.1 Omkoding av variabler

Vi har valgt å kode om kategorivariablene: *par_emp, par_edu, books_home* og *speak_testlang_home* til dummyvariabler. Når man bruker kategorivariabler, gir det ikke alltid mening å tolke effekten av de ulike kategoriene. Ta for eksempel variabelen vår *speak_testlang_home*. Den er delt inn i 3 kategorier ut ifra hvor mye norsk man snakker i hjemmet: 1 = alltid, 2 = noen gang og 3 = aldri. Hvor stort er omfanget av kategori 2 her? gir det mening å legge like stor vekt på en økning fra 1 til 2 som fra 2 til 3? Noen ganger kan det være hensiktsmessig og mer presist å kode om slike variabler. Hovedpoenget med å gjøre en slik omkoding er at det gir mulighet til å se den isolerte effekten av den kategorien/tallkoden som er blitt omgjort til dummy. I stedet for å få et snitt av korrelasjonen til kategorivariabelen med avhengig variabel, velger vi heller å «plukke ut» tallkoder for å få frem korrelasjonen dersom denne kategorien er sann. På denne måten kan man se den isolerte effekten av å alltid snakke norsk i hjemmet sett opp mot å aldri gjøre det, eller en universitetsutdanning sett opp mot ingen utdanning.

Siden kategoriene i en kategorivariabel vil være gjensidig utelukkende lager man én mindre dummy enn man hadde kategorier i kategorivariabelen. Dermed unngår man å havne i «the dummy-variable trap», som går ut på at to dummyvariabler overlapper hverandre. Et eksempel på dette er hvis man hadde hatt to dummyvariabler for kjønn, en som er sann for gutt og en som er sann for jente. Konsekvensen blir da at man får perfekt multikollinearitet.

For å unngå dette utelater man en kategori når man koder om. Dette vil bli referansepunktet til de andre dummyvariablene.

4.2 Presentasjon av data

Tabell 4.1 – *Oversikt over variabler. Alle variabler er kodet om slik at de tar norske navn.*

*Leseresult**	Norske testresultater for lesing. Kontinuerlig variabel, sier hva studenten oppnådde på testresultatene i lesing. Målt på en skala fra 0-700.
utlandForldr	Utenlandske foreldre. Dummyvariabel. 1 = begge foreldrene er fra utlandet, 0 = begge foreldrene er norske.
utlandStud	Norskfødt student. Dummyvariabel. 1 = studenten er ikke er født i Norge, 0 = studenten er født i Norge.
*inntekt**	Inntekt. kategorivariabel som rangerer husholdningens samlede inntekt målt i USD på en skala fra 1-6. 1 = mindre enn $20.000, 6 = mer enn 60.000$.
barnehage	Barnehage. Dummyvariabel som forteller om studenten gikk i barnehage eller ikke. 1 = gikk i barnehage, 0 = gikk ikke i barnehage.
universitet	Foreldre med universitetsgrad. Dummyvariabel som forteller minst en av foreldrene har en universitetsgrad. 0=ingen universitetsgrad, 1=universitetsgrad.
fulltid2	Foreldres jobbsituasjon. Dummyvariabel som forteller om begge foreldrene jobber fulltid eller ikke. 0=begge jobber ikke fulltid, 1=begge jobber fulltid.
fulltid1	Foreldres jobbsituasjon. Dummyvariabel som forteller om en av (og ikke begge) foreldrene jobber fulltid. 0=false, 1= en av (og ikke begge) foreldrene jobber fulltid (true).
deltid	Foreldres jobbsituasjon. Dummyvariabel som forteller om begge foreldrene jobber deltid. 0=1 eller flere av foreldrene jobber fulltid, 1=begge jobber mindre enn fulltid.
NorskHjemme	Snakker norsk i hjemmet. Dummyvariabel som forteller om det snakkes norsk hele tiden i hjemmet. 0=snakker ikke alltid norsk i hjemmet, 1=snakker noe norsk i hjemmet
noeNorsk	Snakker norsk i hjemmet. Dummyvariabel som forteller om det snakkes norsk noen ganger i hjemmet. 0=snakker ikke “noe” norsk i hjemmet, 1=snakker noe norsk i hjemmet

bøker100	Antall bøker i hjemmet. Dummyvariabel som sier noe om hvor mange bøker studenten har tilgjengelig i hjemmet. 0=mindre enn 100 bøker, 1=mer enn 100 bøker

Tabell 4.1 viser hvilke variabler vi ønsker å bruke i modellen vår. *Inntekt* og *leseresult* er markert med (*), som betyr at disse ikke er dummyvariabler. Videre er det viktig å være klar over skalaen variablene er målt på. Dummyvariablene er enten sanne eller usanne, skalaen går fra 0-1. Man vil derfor få frem den isolerte effekten av for eksempel *barnehage* dersom variabelen er sann, sammenlignet med kategorivariabelen *inntekt* hvor man får den gjennomsnittlige effekten av en endring. Senere får vi se at *beta-koeffisienten* til *barnehage* er 9.3, mens *inntekt* har en beta-koeffisient på 4.7. Beta koeffisienten måler hvor mye den avhengige variabelen i snitt endres når vi har en marginal endring i variabelen vi ser på. Om vi ikke tar hensyn til skalaen vil man feilaktig tolke dette som at det å gå i barnehagen har en større sammenheng med leseferdigheter enn familiens samlede inntekt.

4.3 Deskriptiv statistikk til avhengig variabel

Tabell 4.2 – Deskriptiv statistikk for den avhengige variabelen "leseresult"

Variabel	observasjoner	Gjennomsnitt	Standardavvik	Minimum	Maximum
leseresult	3,459	498.2563	78.36616	228.0606	695.8717

Tabell 4.2 viser deskriptiv statistikk over vår avhengige variabel *leseresult.* Denne variabelen viser hvordan norske elever gjør det på PIRLS sin lesetest fra 2001. Som vi kan se beregner den data ut ifra 3 459 observasjoner. Dette er en kontinuerlig variabel. Her er det mulig å oppnå en score mellom 0 og 700. Gjennomsnittet viser alle leseresultatene delt på antall elever, i dette utvalget er gjennomsnittlig resultat 498.2563. Det gjennomsnittlige avviket fra gjennomsnittet er på 78,366, og er standardavviket til utvalget (Thomas, 2005, s.12). Vi kan merke oss at vi har et relativt stort standardavvik og distansen mellom minimumsverdien og maksimumsverdien er stor. Videre i oppgaven vil vi peke på noen av grunnene til at differansen er såpass stor.

4.4 Deskriptiv statistikk for interessevariabel

Tabell 4.3 – Deskriptiv statistikk for interessevariabelen vår "utlandForldr"

Variabel	Observasjoner	Gjennomsnitt	Standardavvik	Minimum	Maximum
utlandForldr	3,374	.0583877	.2345098	0	1

Målet vårt er å undersøke effekten av utenlandske foreldre på leseresultater. Det vil si at *utlandForldr* blir vår interessevariabel og *leseresult* blir vår avhengige. Ovenfor kan man se deskriptiv statistikk for interessevariabelen vår som er en dummyvariabel med 3 374 observasjoner. Gjennomsnittlig verdi her er nesten 0.06. Det vil si at ca. 6% av respondentene har foreldre som ikke er født i Norge.

4.5 Deskriptiv statistikk for kontrollvariablene

Tabell 4.4 – *deskriptiv statistikk for kontrollvariabler*

Variabler	*Observasjoner*	*Gjennomsnitt*	*Standardavvik*	*Minimum*	*Maximum*
utlandStud	3,355	.0909091	.2875226	0	1
barnehage	3,137	.8603762	.3466516	0	1
universitet	3,098	.5429309	.4982339	0	1
fulltid2	2,975	.4376471	.4961803	0	1
fulltid1	2,975	.5152941	.4998501	0	1
deltid	2,975	.010084	.0999285	0	1
bøker100	2,459	.7025152	.4572177	0	1
norskHjemme	3,389	.9023311	.2969103	0	1
noeNorsk	3,389	.0835055	.2766855	0	1
inntekt	2,994	4.064128	1.550755	1	6

Tabell 4.4 viser deskriptiv statistikk for kontrollvariablene våre. Dette er variabler vi velger å inkludere for å "kontrollere" at vi måler sammenhenger mellom den avhengige variabelen *leseresult* og interessevariabelen *utlandForldr*.

4.5.1 Beskrivelse av tabellen

I tabell 4.4 kan man se gjennomsnitt, standardavvik, antall observasjoner og hvilke verdier kontrollvariablene våre kan ta. Det som er viktig å legge merke til her er at alle variablene utenom *inntekt* er dummyvariabler. Tabellen vår viser oss blant annet at 9% av elevene ikke er født i Norge og 54% har foreldre med universitetsgrad. Vi får bare en verdi for standardavviket til dummyvariabler. Det er fordi den bare kan ta to verdier, 0 eller 1. En dummyvariabel hvor objektene er perfekt fordelt mellom 0 og 1 ville derfor hatt et snitt på 0.5 og ett standardavvik på 0.5. Verdiene til kategorivariabelen *inntekt* forteller oss at gjennomsnittsverdien er 4, altså kategorien $40,000-49,999. Det betyr en

gjennomsnittsinntekt per husholdning på 360,000 kr. Dette stemmer godt med gjennomsnittlig husholdningsinntekt fra 2001 (SSB, 2003), som tyder på at utvalget er representativt. Vi ser også at standardavviket er relativt høyt, noe som viser en stor spredning i verdiene våre. Tabellen nedenfor viser fordelingen på de ulike kategoriene for variabelen *inntekt* og bekrefter dette.

Tabell 4.5 – *spredning for kontrollvariabelen inntekt*

inntekt	antall	prosent	kumulativ
1, < $20,000	193	6.45	6.45
2, $20,000-$29,999	370	12.36	18.80
3, $30,000-$39,999	512	17.10	35.91
4, $40,000-49,999	625	20.88	56.78
5, $50,000-$59,999	565	18.87	75.65
6, > $60,000	729	24.35	100
total	**2,994**	**100.0**	

Ved utforming av modellen er det viktig å velge uavhengige variabler som har en årsakssammenheng med den avhengige variabelen. Dette kan forklares ved et eksempel: Haiangrep har for eksempel en korrelasjon med antall is konsumert. Er dette fordi is gjør menneskekroppen mer attraktiv for haier?

Selvfølgelig ikke. Dette eksemplet demonstrerer forskjellen på sammenheng og årsakssammenheng. De fleste haiangrep skjer på strender, samtidig spiser de fleste mer is når de er på stranden en varm sommerdag enn når man er på fjellet. Om man skulle testet sammenhengen mellom is konsumert og haiangrep ville derfor gode kontrollvariabler vært "tid på stranda", "antall timer i havet" og dummyvariabelen "sommer". Derfor er det viktig når man velger kontrollvariabler å tenke godt over hvilke andre faktorer som kan påvirke den avhengige variabelen.

Siden vi har interessevariabelen "utenlandske foreldre" har vi valgt kontrollvariabler som er korrelert med "utenlandske foreldre" og har effekt på leseresultater. På denne måten blir det lettere å finne ut om *utlandForldr* er en "is" eller "bading", eller muligens en kombinasjon.

Vi ser for oss at faktorer som påvirker hjemmet kan ha direkte innvirkning på leseresultater. Innvandrere i Norge tjener i snitt mindre enn landsgjennomsnittet, de er også overrepresentert på arbeidsledighetsstatistikken. At innvandrere generelt tjener mindre og har høyere ledighet er ikke særlig omstridt (SSB, 2020). Men det kan være noe kontrovers i grunnene til dette, for eksempel integreringsproblemer. På grunn av dette har vi valgt å inkludere variabelen *inntekt*. Resterende variabler begrunnes på grunnlag av lignende argumentasjon for samfunnsaspekter (Grønmo, 2014).

4.6 Presentere matrise med enkel korrelasjon mellom variablene

Tabell 4.6 – *Korrelasjonsmatrise som viser korrelasjonen mellom alle variablene våre*

	leseresult	*utlandForldr*	*utlandStud*	*barnehage*	*universitet*	*fulltid2*	*fulltid1*	*deltid*	*books100plus*	*norskHjemme*	*noeNorsk*	*inntekt*
leseresult	1											
utlandForldr	-.1303	1										
utlandStud	-.1169	.2600	1									
barnehage	.0781	-.0250	.0415	1								
universitet	.2915	-.0385	.0114	.1355	1							
fulltid2	.0646	-.0385	-.0160	.1384	.1471	1						
fulltid1	-.0310	.008	-.0014	-.1246	-.106	-.9136	1					
deltid	-.0807	.0818	.0668	-.0708	-.0462	-.0899	-.1022	1				
books100plus	.2466	-.1510	-.0472	.0443	.3577	.0938	-.0662	-.0368	1			
norskHjemme	.14477	-.4355	-.1904	-.0053	.0361	.0010	.0338	-.0782	.1146	1		
noeNorsk	-.1060	.3518	.1402	.0323	-.0248	.0089	-.0324	-.0440	-.0840	-.9299	1	
inntekt	.2152	-.1320	-.0451	.1099	.3896	.3345	-.2331	-.1212	.2969	.1063	-.0896	1

Når vi leser korrelasjonsmatrisen, er vi interessert i korrelasjonen mellom kontrollvariablene og den avhengige og uavhengige variabelen. Dette er en fin måte å sjekke for om det er valgt gode kontrollvariabler. Som forklart i del 4.6 ønsker vi å inkludere kontrollvariabler for å utelukke sammenhengen mellom den uavhengige (interessevariabel) og avhengige variabelen.

Hva forteller korrelasjonsmatrisen oss? Det første vi kan merke oss er at det ikke er noen av kontrollvariablene våre som er positivt eller negativt korrelert med både *leseresult* og

utlandForldr. De som er positivt korrelert med *leseresult* er negativt korrelert med *utlandForldr* og omvendt. Dette tyder på at kontrollvariablene vi har valgt ut er verdt å kontrollere for.

Variablene *inntekt, bøker100* og *utlandStud* er alle høyt korrelert med både *leseresult* og *utlandForldr*. Det betyr at de har en relativt sterk sammenheng med både vår avhengige og uavhengige variabel. De blir derfor viktige variabler å ha meg seg videre i analysen. Hadde en av disse blitt utelatt hadde sannsynligheten for å konkludere feil grunnet det vi kaller "omitted variable bias" (Studenmund, 2017, s.158) vært relativt stor. Derimot er *universitet* sterkt positivt korrelert med leseresultater, men ikke med utenlandske foreldre. Man kan derfor sannsynligvis si at studenter som har foreldre med universitetsgrad gjør det bedre på lesetester. Det ville blitt feil å si at studenter med utenlandske foreldre gjør det dårligere siden deres foreldre ikke har universitetsgrad. Videre har de omkodede dummyvariablene *fulltid2* og *fulltid1* en korrelasjon med hverandre på -0.9136, mens *norskHjemme* og *noeNorsk* har en korrelasjon med hverandre på -0.9299. Dette er en indikator på multikollinearitet, som kan gjøre det vanskelig å skille mellom den isolerte effekten til variablene. I kapittel 5 kommer vi tilbake til hvordan vi løser dette.

4.7 Styrker og svakheter med datasettet

Dataene vi bruker kommer fra PIRLS og består av over 3000 observasjoner. Siden vi har såpass mange observasjoner vil dataene våre være svært robuste. Dette er en anerkjent og god datakilde som har gitt liv til flere ulike rapporter og artikler. Siden vi bruker en åpen datakilde blir det lett for andre å etterprøve resultatene, noe som er svært viktig. Dog er de vesentlige trekkene i analysen vår lik. Siden funnene våre samsvarer med SSBs rapporter fra 2016 tyder det på at resultatene fortsatt er relevante (Dzamarija, 2017).

Som nevnt i kapittel 2 består skoleproduktfunksjonen av karakteristika ved skolene, lærere og pensum i tillegg til familie- og medelevkarakteristika. En mulighet for oss kunne vært å inkludere en modell der vi også kontrollerer for ulike aspekter ved skolene. Ved å gjøre dette ville vi fått et mer nyansert resultat med tanke på for eksempel demografi. Det faktum at en skole i Oslo sannsynligvis har en høyere andel innvandrere og norskfødte med innvandrerforeldre enn en skole i Finnmark vil kunne være med å påvirke den avhengige variabelen *leseresult.* Men en slik utvidelse ville ført til at lengden på besvarelsen gikk langt utover de rammene som er fastsatt hva gjelder antall sider. Av praktiske hensyn blir vi nødt til å begrense omfanget, og velger derfor å kun inkludere familiekarakteristika i vår modell.

4.8 Oppsummering

For å kunne se den isolerte effekten av kategorivariablene, har vi omgjort flere av dem til dummyer. Kapittelet har tatt for seg deskriptiv statistikk for avhengig variabel der man kunne se at leseresultatene til elevene som dataen baserer seg på kunne være mellom 0 og 700. I den deskriptive statistikken for kontrollvariabelen *utlandForldr*, kunne man se at 6% av elevene har foreldre som ikke er født i Norge. Deretter kommenterte vi tabellen med deskriptiv statistikk for kontrollvariablene. Kontrollvariablene er de variablene vi tror kan ha en sammenheng med interessevarabelen vår. For å se på korrelasjonen mellom kontrollvariablene og den avhengige og de uavhengige variablene ble det presentert og tolket en korrelasjonsmatrise. Til slutt ble det tatt opp som kritikk, at dataen vi bruker kommer fra 2001, som er 19 år siden. Det er rimelig å anta at enkelte forhold har endret seg siden det, men ut ifra andre studier virker de vesentlige trekkene i analysen å ikke være særlig påvirket.

5. Regresjonsanalyse av empiriske resultater

I dette kapittelet skal vi utføre regresjonsanalysen, som vil si å bruke minste kvadraters metode for å analysere datasettet. Først argumenterer vi for hvorfor vi benytter en multivariabel lineær funksjon, deretter gjennomfører vi regresjon på en restriktiv modell som kun inneholder avhengig variabel og interessevariabelen vår. Videre utvider vi modellen ved å legge til kontrollvariablene. Intuisjonen bak en slik utvidelse er at noe så sammensatt som barns læring ikke bare avhenger av hvorvidt foreldrene er utenlandske eller ikke. Resultater fra relevant forskning presentert i kapittel 2 viser også at det er andre faktorer som gir utslag på leseresultatene i tillegg til vår interessevariabel. Til slutt gjennomfører vi en regresjonsanalyse på vår utvidede problemstilling, om jenter gjør det bedre enn gutter blant barn av utenlandske foreldre.

5.1 Valg av funksjonsform

Det er forutsetningen for minste kvadraters metode er at vi har en lineær sammenheng mellom de avhengige og uavhengige variablene. Det betyr ikke at den underliggende teorien må være lineær. Det er fullt mulig å korrigere for ikke lineære sammenhenger ved å ta ulike matematiske grep. Man kan for eksempel sette opp en log funksjon eller kvadrere ulike ledd av funksjonen. Likevel så vil vi i vårt tilfelle ta utgangspunkt i en vanlig multivariabel lineær funksjon.

$$Y = \alpha + \Sigma\, \beta_j\, X_j + \varepsilon_i$$

I funksjonen over er Y vår avhengige variabel, mens X_j er våre uavhengige, α (alfa) er konstantleddet og ε_i (epsilon) er restleddet vårt.

Grunnen til at vi velger å ta i bruk en multivariabel lineær funksjon er fordi vi nesten utelukkende har dummyvariabler. Siden dummyvariabler kun kan ta to verdier, og en linje mellom to verdier alltid vil være lineær, blir det naturlig å velge en vanlig lineær funksjon.

5.2 Presentasjon av modellen og enkel regresjon

Vår oppgave vil benytte seg av en skoleproduktfunksjon og vi bruker lesetestresultater som den avhengige variabelen, eller Y i modellen vår.

Det vil bli presentert 4 modeller. Modell I er en svært restriktiv modell som kun ser på den avhengige variabelen og interessevariabelen vår. Her ser vi kun på sammenhengen mellom testresultater og det å ha utenlandske foreldre. Dette er en skoleproduktfunksjon fordi den ser på virkingen på testresultater. I denne modellen kontrollerer vi ikke for noen andre faktorer. Vi bruker ikke denne modellen fordi vi tror at den vil ha den beste forklaringskraften, men vi velger å ha den med for å demonstrere viktigheten av å kontrollere for relevante faktorer. Altså viktigheten av å unngå å utelatte relevante variabler slik at man unngår "omitted variable bias".

Modell I: *leseresult* = α + β *utlandForldr* + epsilon

Basert på intuisjon og tidligere forskning så forventer vi at *utlandForldr* vil ha en negativ sammenheng med leseresultater. For å teste dette utfører vi en MKM-regresjonsanalyse i Stata for å estimere betakoeffisienten, β, konstantleddet, standardavvik, og t-verdi. Vi kan bruke disse verdiene til å predikere Y-verdien. Når vi anvender MKM-analyse blir støyleddet, epsilon, lik 0 (Thomas, 2005, s. 359). Derfor behøver vi ikke ta hensyn til støyleddet når vi estimerer modellen. Resultatene fra regresjonen er vist i tabell 5.1.

leseresult = 502.774 + (-58.433* *utlandForldr*)

Modellen forteller oss at de studentene med utenlandske foreldre (*utlandetForldr* =1) sin estimerte lesetestscore er 444.341, altså 58.433 poeng mindre enn de med norske foreldre. Den negative endringen på 58.433 er beta-koeffisienten når *utlandForldr* er sann. Konstantleddet er den estimerte verdien der alle uavhengige variabler er lik sin laveste verdi. Her blir det lik estimert lesetestresultat for barn med norske foreldre (*utlandForldr* =0). Disse

resultatene er ikke overaskende og samsvarer med forventningene våre. Vi ønsker å teste om resultatene våre er signifikante, og setter derfor opp en t-test med hypotesene:

H_0: Det **er ikke** en negativ sammenheng mellom *utlandForldr* og *leseresult*, $\beta = 0$.

H_A: Det **er** en negativ sammenheng mellom *utlandForldr* og *leseresult*, $\beta < 0$

Vi velger ett signifikansnivå på 0.05. Det betyr at vi godtar en 5% sannsynlighet for at vi begår en type I feil. Ved hjelp av t-fordelingstabellen til Thomas i boken "using statistics in economics", finner vi at kritisk verdi for testen vår er 1.64 (Thomas, 2005, s. 587). Vi finner kritisk verdi (TS) ved:

$$TS=b_j/S_{bj}$$

Dersom absoluttverdien til TS er større enn kritisk verdi, forkaster vi nullhypotesen.

$$TS= -58.433/ 5.642987 = -10.35$$

$$|-10.35|>1.64$$

Som vi også kan lese ut ifra tabell 5.1 gir utregningen oss en t-verdi på -10.35, hvor absoluttverdien er større enn den kritiske verdien på 1.64. Kan konkludere med at sammenhengen mellom *leseresult* og *utlandForldr* er signifikant til et 5% signifikansnivå. Dermed har vi grunnlag for å si at leseferdighetene til elevene påvirkes av om foreldrene deres er født i utlandet eller ikke, og vi kan dermed forkaste H_0. Det er likevel viktig å poengtere at dette er en svært restriktiv modell. I denne modellen er justert $R^2 = 0.0305$, det vil si at det å ha utenlandske foreldre kun forklarer 3.05% av variasjonen i leseresultater. Innenfor et 95 % konfidensintervall har vi et spenn på beta-koeffisienten mellom –69.497 og –47.369, noe som betyr at 95% av barna med utenlandske foreldre gjør det mellom –47.369 og –69.497 poeng dårligere enn barn med norske foreldre.

Tabell 5.1 – *Enkel regresjonsanalyse for leseresult, utlandetForldr*

Variabler	*Betakoeffisien*	*Standardavvik*	*t-verdi*	*95% konf. intervall*
utlandForldr	-58.433***	5.642987	-10.35	-69.497 , -47.369
Konstant	502.774***	1.364	368.73	

**** uttrykker at variabelen er signifikant ved et 1% signifikansnivå, **5%, *10%*

5.3 Utvidet modell

Modell I fant at effekten av *utlandForldr* var signifikant, men som vist var modellen svært restriktiv. I modell I ser vi kun på foreldres fødested og leseresultater uten å kontrollere for noen andre faktorer som kan ha en påvirkning. Dette er noe snevert når vi ser på et tema som er så sammensatt som barns læring. Som nevnt i kapittel 3.2.2 og 4.7 ønsker vi å kontrollere for underliggende faktorer som kan ha påvirket resultatene våre i modell I. Vi utvider modellen til å kontrollere for familie- og elevkarakteristika så vi kan analysere den isolerte effekten av *utlandForldr*.

Faktorene i modell II er helt eller delvis utenfor myndighetenes kontroll. Vi bruker fortsatt en lineær skoleproduktfunksjon, nå med 8 dummyvariabler og en kategorivariabel. Tabell 5.2 viser en oversikt over alle de nye variablene vi har lagt til i modell II sammen med resultatene. Modellen kan vises som denne ligningen:

$$leseresult = \alpha + \gamma_1\, utlandForldr + \gamma_2\, utlandStud + \gamma_3\, barnehage + \gamma_4\, universitet + \gamma_5\, fulltid2 + \gamma_6\, fulltid1 + \gamma_7\, deltid + \gamma_8\, bøker100 + \beta_9\, inntekt$$

For å estimere modellen fyller vi inn for betakoeffisientene fra tabell 5.2. Både γ og β står for beta-koeffisient, mens γ brukes foran dummyvariabler.

Vi har valgt å ta med p-verdien i tabellen for å vise laveste signifikans som H0 kan forkastes ved.

Tabell 5.2 – *regresjonsanalyse for modell II*

Variabler	***Beta-koeffisient***	***Standardavvik***	***t-verdi***	***P-verdi***
utlandForldr	-22.971	6.74	-3.41	0.001
utlandStud	-26.016	5.264	-4.94	0.000
barnehage	8.507	4.134	2.06	0.040
universitet	32.125	3.129	10.27	0.000
fulltid2	-0.520	7.888	-0,07	0.947
fulltid1	2.538	7.666	0.33	0.741
deltid	-31.833	15.646	-2.03	0.042
bøker100	22.332	3.242	6.89	0.000
inntekt	3.870	1.062	3.65	0.000

konstant	450.642	8.574	52.56	0.000
Observasjoner	**2,734**	***Adj R-kvadrert***	**0.1349**	

Vi kan estimere lesetestresultatene til de fleste typer studenter ved hjelp av modellen vår. Alt man behøver å gjøre da er å fylle inn for de verdiene man ønsker å se på. Om vi vil se på student med foreldre med universitetsutdanning og tilgang til 100 bøker i hjemmet så fyller man inn for disse verdiene i ligningen. De variablene vi ikke ønsker å se på setter vi lik 0. Det som er viktig å tenke på da er hva verdien 0 innebærer. I dette tilfellet betyr det for eksempel at vi ser på en student som aldri har gått i barnehage, er født i Norge, har norske foreldre, foreldrene er arbeidsløse og de har en samlet inntekt på under $20.000 i året. Denne eleven har en estimert testscore på 505, altså marginalt bedre enn gjennomsnittet på 498.

Grunnet vår problemstilling er vi hovedsakelig ute etter å finne effekten av utenlandske foreldre. Vi kan derfor holde de andre variablene konstant og fokusere på γ_1, beta-koeffisienten til *utlandForldr*. Vi ser i tabell 5.2 at γ_1 er -22.971. γ_1 har blitt redusert kraftig fra modell I til modell II fordi vi i modell II kontrollerer for flere variabler som forklarer endringen i *leseresult.* Effekten av disse variablene ble sannsynligvis "overført" til *utlandForldr* i modell I. For å teste om resultatene våre er signifikante setter vi opp en hypotesetest:

H_0: $\gamma_1 \geq 0$, Det **er ikke** en negativ sammenheng mellom leseresultater og utenlandske foreldre.

H_A: $\gamma_1 < 0$, Det **er** en negativ sammenheng mellom leseresultater og utenlandske foreldre

Vi velger igjen et signifikansnivå på 5% og siden vi har en ensidig test gir dette oss en kritisk verdi på 1.64. Vi har en t-verdi på -3.41. Siden absoluttverdien av t-verdien vår er større en kritisk verdi kan vi forkaste H_0. Det er også verdt å legge merke til at H_A er signifikant selv på 1% signifikansnivå i både modell I og II. P-verdien viser laveste signifikansnivå man kan forkaste nullhypotesen på. Fra tabell 5.2 har *utlandForldr* en p-verdi på 0.001, kan altså forkaste nullhypotesen på et 0.1 % signifikansnivå.

Som nevnt i kapittel 4.7 under korrelasjonsmatrisen har vi en sterk korrelasjon mellom de omkodede variablene *fulltid2* og *fulltid1*. Regresjonen i kapittel 5.3 viser at begge variablene har svak korrelasjon med *leseresult,* en korrelasjon som ikke er signifikant på et 5%

signifikansnivå. Derfor vil det være formålstjenlig å teste hvorvidt disse variablene sammen med dummyen *deltid*, som heller ikke er signifikant, har en effekt på *leseresult*. Kjører derfor en f-test, for å undersøke nullhypotesen om at de tre variablene *fulltid2*, *fulltid1* og *deltid* ikke har noen innvirkning på *leseresult.* Beholder nullhypotesen med en f-verdi på 2.27 med 3 frihetsgrader i teller og 2724 frihetsgrader i nevner. Siden vi ikke kan forkaste nullhypotesen kan vi heller ikke fastslå at foreldrenes arbeidssituasjon har noen signifikant effekt på leseresultater.

5.4 Modell III

$$leseresult = \alpha + \gamma\_1\ utlandForldr + \gamma\_2utlandetStud + \gamma\_3barnehage + \gamma\_4universitet + \gamma\_5fulltid2 + \gamma\_6fulltid1 + \gamma\_7deltid + \gamma\_8bøker100 + \beta\_9inntekt + \gamma\_10norskHjemme + \gamma\_11noeNorsk$$

Tabell 5.3 Modell III regresjon

leseresult	*beta-koeffisient*	*Standardavvik*	*t-verdi*	*p-verdi*
utlandForldr	-7.731	7.391	-1.05	0.296
utlandStud	-22.938	5.286	-4.34	0.000
barnehage	7.893	4.135	1.91	0.056
universitet	32.120	3.130	10.26	0.000
fulltid2	-1.241	7.884	-0.16	0.875
fulltid1	1.125	7.665	0.15	0.883
deltid	-28.110	15.602	-1.80	0.072
bøker100	21.116	3.247	6.50	0.000
norskHjemme	63.527	14.417	4.41	0.000
noeNorsk	43.909	14.691	2.99	0.003
inntekt	3.778	1.063	3.55	0.000
konstantleddet	391.127	16.288	24.01	0.000
observasjoner	**2,713**	***adj R-kvardrert***	**0.1397**	

På samme måte som for *fulltid2*, *fulltid1* og *deltid* kjører vi en f-test på *noeNorsk* og *norskHjemme*. Nullhypotesen om ingen sammenheng mellom de to variablene og *leseresult* forkastes både på 5% og 1% signifikansnivå. F-verdien er på 13.58 med 2 frihetsgrader i teller og 2701 frihetsgrader i nevner. De høye beta-koeffisientene for *norskHjemme* og

noeNorsk har en signifikant effekt på *leseresult.* Som nevnt i kapittel 4.7 kan den høye korrelasjonen mellom *fulltid2* og *fulltid1* samt *noeNorsk* og *norskHjemme* tyde på multikollinearitet. Dersom man har perfekt korrelasjon mellom to uavhengige variabler, vil dette være et brudd på forutsetning VI for minste kvadraters metode. Vi velger å ikke gjøre noen endringer på modellen for å unngå "omitted variable bias", men tar med følgende funn videre: Foreldrenes arbeidssituasjon har ikke noen signifikant effekt på leseresultater, mens *noeNorsk* og *norskHjemme* har en signifikant effekt på *leseresult.*

De to nye variablene øker ikke forklaringskraften til modellen (adj R-kvadrert) noe betydelig, men introduksjonen av dem reduserer betakoeffisienten til *utlandForldr* betraktelig. Dette kan tolkes som at *utlandForldr* i seg selv ikke er tilstrekkelig for å forklare variansen i *leseresult,* og at grunnen til at vi så en stor effekt i *utlandForldr* tidligere skyldes hvilket språk man prater i husholdningen. I tillegg er det rimelig å anta at det er vanligere for familier med utenlandske foreldre at man ikke snakker norsk i hjemmet. Vi kan se effekten av dette i korrelasjonsmatrisen (tabell 4.6).

For å teste om vi fortsatt kan forkaste H_0 setter vi igjen opp hypotestesten vår. Vi bruker fortsatt en ensidig test med et signifikansnivå på 5%. Kritisk verdi blir fortsatt |1.64|.

> H_0: $\gamma_1 \geq 0$, Det **er ikke** en negativ sammenheng mellom leseresultater og utenlandske foreldre.
>
> H_A: $\gamma_1 < 0$, Det **er** en negativ sammenheng mellom leseresultater og utenlandske foreldre.

I modell III har *utlandForldr* en t-verdi på -1.05. Siden 1.05 er mindre enn 1.64 kan vi ikke forkaste H_0. Det er ikke en signifikant negativ sammenheng mellom leseresultater og utenlandske foreldre.

Selv om resultatet kan virke noe overaskende, stemmer dette godt overens med konklusjonen til Coleman i 1966 som sier "It's all in the family" (Bonesrønning, 2004, s. 16). Testresultatene har altså ingen årsakssammenheng med utenlandske foreldre, her er det andre faktorer som spiller inn. Ut ifra tabell 5.3 er de viktigste faktorene språket man snakker hjemme, foreldres utdanning, eget fødested og antall bøker i hjemmet. Det ser ut til at foreldrenes arbeidssituasjon også kan være viktig, men denne sammenhengen er ikke signifikant på 5% nivået.

5.5 Tolkning av resultater

Modell I og II forteller oss at vi kan forkaste H_0, og dermed godta at det er en negativ sammenheng mellom det å ha utenlandske foreldre og lesetestresultater. Modell III derimot forteller oss at vi ikke har tilstrekkelig grunnlag for å forkaste H_0. Modell III er den mest omfattende modellen vi har, der vi har kontrollert for alle variabler vi har sett tidligere litteratur legger vekt på, utenom kjønn. Vi har ikke kontrollert for kjønn fordi det ikke er en variabel som har noen sammenheng med familier med utenlandske foreldre. Modell II er også ganske omfattende, men kontrollerer ikke for språket som brukes i hjemmet. I modell I kontrolleres det ikke for noen ting.

Når vi analyserer verdiene i regresjonsanalysene våre, ser vi at jo flere kontrollvariabler vi legger til, jo mer synker betakoeffisienten til *utlandForldr,* og jo lavere t-verdi får den. Den får altså en mindre effekt, og blir mindre signifikant. Ved å bruke modell III, kan vi konkludere med at vi ikke har tilstrekkelig grunnlag til å si at vi har en negativ sammenheng mellom *utlandForldr* og *leseresult*. Konklusjonen er derfor at vi ikke kan forkaste nullhypotesen vår. Dette betyr ikke at modell I og II vil være dårlige til å predikere hvordan elever med utenlandske foreldre gjør det på lesetester.

Som modell I viser, kan vi se at 95% av de elevene med utenlandske foreldre får mellom 69.497 og 47.369 mindre poeng enn de elevene med norske foreldre. Det som er viktig å understreke her, er at dette ikke skyldes at foreldrene ikke er født i Norge, men det skyldes andre underliggende grunner som har en sammenheng med variabelen *utlandForldr*.

Foreldre som er født i utlandet sender barna sine i mindre grad i barnehagen, færre har universitetsutdanning, de jobber mer deltid, de har lavere inntekt, de prater mindre norsk i hjemmet og det er større sannsynlighet for at deres egne barn heller ikke er født i Norge. Modell III viser oss at det er på grunn av disse faktorene at barna deres gjør det dårligere på lesetestene. De fleste av disse grunnene har ikke noe med den enkelte forelder å gjøre.

Det er mye debatt om grunnene til hvorfor det er sånn, men det er ikke kontroversielt å si at et integreringssystem som ikke er godt nok, og en systematisk diskriminering kan være noe av grunnen. Systemendringer som å redusere barnefattigdom, øke andelen tilgjengelige bøker og bedre integreringen kan derfor være potensielt gode tiltak for å øke leseferdighetene hos barn og unge. Dette er spennende og vanskelige problemstillinger. Likeså vil ikke denne

oppgaven gå videre inn på dem, da vår hensikt er å forklare om det er en statistisk signifikant sammenheng mellom leseferdigheter og studenter med foreldre som er født utenfor Norge.

Når vi anvender modell III så ser vi at vi ikke har tilstrekkelig grunnlag til å forkaste nullhypotesen vår om at det er en negativ sammenheng mellom å ha foreldre som er født utenfor Norge og leseferdigheter. Derimot så finner vi at en av hovedgrunnene til at elever gjør det dårligere på lesetesten er at de ikke blir eksponert nok for norsk i hjemmet. Det kan virke som at dette skillet kan virke litt kunstig da det er mye flere med utenlandske foreldre som ikke snakker norsk i hjemmet enn i hjem med norske foreldre. Men her er det viktig å skille mellom en sammenheng og en årsakssammenheng. Andre viktige faktorer som forklarer variansen i leseresultatene er om foreldrene har universitetsutdanning, hvor mange bøker de har tilgjengelig i hjemmet og husholdningens samlede inntekt.

Vi kan ikke konkludere med at det er signifikant negativ sammenheng mellom variabelen *utlandForldr* og *leseresult.* Likevel samsvarer funnene våre med funnene til Coleman, Dzamarija og Bøyesen (Bøyesen, 2004; Coleman, 1966; Dzamarija 2014). Det finnes artikler og undersøkelser som konkluderer med at vi har en sammenheng og det finnes artikler som konkluderer det motsatte. Derfor kan det kanskje virke som at det er stor splid på feltet i litteraturen, dette stemmer ikke. De fleste seriøse aktører er enige om at innvandrere og norskfødte med utenlandske foreldre i snitt gjør det dårligere enn norskfødte med norske foreldre i lesing. Om man kan trekke en signifikant sammenheng mellom de to kommer an på hvor mange andre faktorer man velger å kontrollere for. Dette vises frem gjennom modellene våre hvor I og II viser en signifikant sammenheng mens modell III ikke gjør det.

At ulike aktører anvender forskjellige modeller skyldes ikke nødvendigvis at de er uenige om hvilke som er best, eller hvilken som blir mest riktig å anvende. At ulike seriøse aktører bruker forskjellige modeller skyldes hovedsakelig hva man ønsker å finne ut av eller hvilken funksjon arbeidet har. På grunn av samfunnsoppdraget til SSB for eksempel så vil det oftere være mer interessant å se forskjellen i snittscoren på tvers av ulike demografiske grupper enn å finne ut om vi kan se på om vi har en faktisk årsakssammenheng mellom to variabler. I vårt tilfelle er alle modellene viktige. Det er viktig å forstå at barn med utenlandske foreldre har et dårligere utgangspunkt for å lese norsk slik at man kan tilrettelegge bedre i skolene, men det er også viktig å vite hvorfor denne gruppen presterer dårligere slik at man vet hvordan man skal tilrettelegge på en god måte.

5.6 Kjøre regresjonsanalyse på vår utvidede problemstilling

At jenter gjør det bedre enn gutter på lesetester er godt dokumentert (Borgonovi et al., 2018, s. 106). Det vi ønsker å finne ut er om effekten av å ha utenlandske foreldre er større for gutter enn for jenter. Derfor har vi generert en ny variabel som heter *jenteXutlandForldr*, som er en interaksjonsvariabel. Denne variabelen tester om vi har samspill mellom variabelen *jente* (0=gutt, 1=jente) og *utlandForldr*. Om det er slik at effekten av utenlandske foreldre er større for gutter enn for jenter så vil beta-koeffisienten til *jenteXutlandForldr* være negativ. Vi setter opp en ensidig hypotesetest med ett 5% signifikansnivå:

H_0: $\beta \geq 0$ Effekten av *utlandForldr* er **ikke mindre** for jenter enn for gutter

H_A: $\beta < 0$ Effekten av *utlandForldr* **er mindre** for jenter enn for gutter.

Vi setter variablene *jente* og *jenteXutlandForldr* inn i modellene våre. Den eneste variabelen vi er interessert i her er *jenteXutlandForldr.* Tabell 5.4 viser en oversikt over alle verdiene vi trenger for teste om effekten av utenlandske foreldre er større for gutter enn for jenter.

Tabell 5.4 – *Oversikt over beta-koeffisient, standardavvik og t-verdi til "jenteXutlandForldr" når vi har lagt til "jenteXutlandForldr" og "jente" i modell I, II og III.*

leseresult	beta-koeffisient	standardavvik	t-verdi
modell I	-3.107	11.335	-0.27
modell II	-2.559	12.921	-0.20
modell III	0.878	12.927	0.07

Igjen så har vi en ensidig test med et 5% signifikansnivå og en kritisk verdi på 1.64. Vi ser fra tabell 5.4 at t-verdien til samspillvariabelen *jenteXutlandForldr* ikke er tilstrekkelig stor i noen av modellene. Det er derfor ikke belegg for å forkaste H_0, og det er ikke statistisk grunnlag til å si at effekten av å ha utenlandske foreldre er større for gutter enn den er for jenter. Dette funnet er robust på tvers av modellene våre som kontrollerer for flere variabler.

6. Oppsummering og konklusjon

Vi har brukt en skoleproduktfunksjon for å analysere effekten av lesetestresultater. Denne kan spores tilbake til Coleman (1966), og siden den gang har den blitt brukt flere ganger for å analysere mange ulike problemstillinger. I dag er det å analysere testresultater for å måle

nivået på skolene noe vi tar for gitt. Derfor gjennomfører vi blant annet nasjonale prøver årlig.

På noen fagfelt gjør norskfødte med innvandrerforeldre det bedre enn gjennomsnittet. Selv om de gjør det dårligere på VGS er det flere andregenerasjonsinnvandrere som gjennomfører høyere utdanning enn landsgjennomsnittet. Likevel så er det konsensus i faglitteraturen om at både innvandrere og norskfødte med innvandrerforeldre er dårligere norsklesere enn snittet (Bøyesen, 2004; Dzamarija, 2016; Lervåg & Melby Lervåg, 2009; Steinkellner, 2017). Dette kan vi også se tydelig ut ifra resultatene våre. Modell I, som kun ser på sammenhengen mellom *utlandForldr* og *leseresult,* viser en klar signifikant sammenheng. Den sier også at 95% av studentene med utenlandske foreldre i snitt oppnår 47-69 poeng mindre enn norskfødte barn med norske foreldre.

Likevel så ser vi i modell III at variabelen *utlandForldr* ikke lenger er signifikant. Dette burde ikke komme som en overraskelse. De færreste tror at andregenerasjonsinnvandrere er dårligere lesere bare fordi de er andregenerasjonsinnvandrere. Det blir nok feil å tenke at forfatterne og forskerne som konkluderer med at norskfødte barn med innvandrerforeldre er dårligere lesere tror det er fordi de har innvandrerforeldre. De rapporterer dette fordi de ser at det er en tendens i skolene våre. Men det er ingen seriøse aktører som trekker en negativ kausal sammenheng mellom å ha innvandrerforeldre og leseresultater, i alle fall ikke offentlig.

Resultatene våre fra modell III burde derfor ikke komme som noen overraskelse. Etter at vi har kontrollert for de forhold som faglitteraturen peker på og som er relevante for vår problemstilling, så ser vi at effekten av *utlandForldr* ikke lenger er signifikant. Vi ser også på modellene våre at jo flere variabler vi kontrollerer for, jo mindre blir effekten av *utlandForldr*. Det betyr at variansen i leseresultatene ikke direkte skyldes *utlandForldr,* men at det skyldes andre sosioøkonomiske forhold som inntekt, jobb, utdanning, språk, barnehage og tilgang på bøker. Det kan være flere årsaker til at studenter med innvandrerforeldre kommer "dårligere" ut på disse parameterne, men effekten blir at de i snitt gjør det dårligere på norske lesetester.

Som vi ser i avsnittet over så er problemstillingen vår noe kompleks, men det er ikke hypotesetesten vi satte opp i starten av oppgaven. Den sier tydelig at vi kun kan forkaste nullhypotesen "Det e**r ikke** en negativ korrelasjon mellom *leseresult* og *utlandForldr"*

dersom $\beta < 0$ og t-verdien er større en |1.64|. Verdiene våre fra modell III oppfyller ikke disse kravene og vi kan derfor ikke forkaste H0. Vi har ikke noe grunnlag for å si at det er en statistisk signifikant negativ sammenheng mellom det å ha innvandrerforeldre og leseresultater. Modell I og II viser en statistisk signifikant sammenheng, men Modell III viser at vi i I og II mangler noen viktige kontrollvariabler og lider av "omitted variable bias".

Jenter er i snitt bedre lesere en gutter, og gutter er i tillegg mer utsatt for miljøendringer enn det jenter er (Borgonovi et al., 2018, s. 106). Det gjør at gutter i større grad står for andelen av elever som skårer dårligst på lesetester. Når det er sagt så gir ikke våre data noe grunnlag for å si at effekten av utenlandske foreldre er større for gutter enn for jenter. Vi kan altså ikke forkaste nullhypotesen vår om at effekten av utenlandske foreldre er lik for begge kjønn. Dette resultatet gjør seg gjeldende på tvers av alle modellene våre.

Om vi hadde hatt en lengre og mer omfattende artikkel så hadde det også vært spennende å se på ulike forhold knyttet opp mot skolen. Vi vet at vi ikke har en lik demografisk fordeling på skolene våre. Det er altså noen skoler med en mye høyere andel første- og andregenerasjonsinnvandrere. I en videre studie kunne det vært interessant å analysere dette nærmere.

7. litteraturliste

Bonesrønning, H. (2004). Utforming av utdanningspolitikken

– Hva kan økonomene bidra med? *Samfunnsøkonomene*, 14-23.

Borgonovi, F., Ferrara, A., & Maghnouj, S. (2018). The gender gap in educational outcomes in Norway.

Bøyesen, A. L. (2004). Begynneropplæring i lesing og skriving for minoritetselever. *Cappenlens NOA-nett. Nettsted for lærere i norsk som andrespråk*.

Camilla Stoltenberg, H. M. A., Rahman Akhtar Chaudhry, Ingrid Fylling, Rune Hausstätter, Mats A. Kirkebirkeland, Arne Lervåg, Katrine Vellesen Løken, Mats Monsen, Camilla Trud Nereid, Terje Ogden. (2019). *Nye sjanser – bedre læring*. NOU - Norges Offentlige Utledninger. Hentet fra https://www.regjeringen.no/contentassets/8b06e9565c9e403497cc79b9fdf5e177/no/pdfs/nou201920190003000dddpdfs.pdf

Coleman, J. S. (1968). Equality of educational opportunity. *Integrated Education*, *6*(5), 19-28.

Dickinson, E. E. (2016). COLEMAN REPORT SET THE STANDARD FOR THE STUDY OF PUBLIC EDUCATION: Johns Hopkins Magazine.

Dzamarija, M. T. (2016). Barn og unge voksne med innvandrerbakgrunn: Demografi, utdanning og inntekt.

Falch, T., Johannesen, A. B., & Strøm, B. (2009). *KOSTNADER AV FRAFALL I*

VIDEREGÅENDE OPPLÆRING Trondheim: Senter for økonomisk forskning AS Hentet fra https://www.regjeringen.no/globalassets/upload/kd/vedlegg/grunnskole/frafall/kostnader-av-frafall.pdf

familiedepartementet, B. o. (2000). *FNs konvensjon om barnets rettigheter*. Regjeringen.no. Hentet fra https://www.regjeringen.no/no/dokumenter/barnekonvensjonen-kortversjon-norsk/id87582/

Frøslie, K. F. (2017). metaanalyse. *Store Norske Leksikon*.

Gabrielsen, N. N., & Lundetræ, K. (2013). Ordavkoding og leseferdighet i PIRLS. *Over kneiken?*, 117.

Grønmo, S. (2014). kontrollvariabel. *Store Norske Leksikon*.

Hanushek, E. A. (2020). Education production functions *The Economics of Education* (s. 161-170): Elsevier.

Kotrlik, J., & Higgins, C. (2001). Organizational research: Determining appropriate sample size in survey research appropriate sample size in survey research. *Information technology, learning, and performance journal*, *19*(1), 43.

Melby-Lervåg, M., & Lervåg, A. (2009). Muntlig språk, ordavkoding og leseforståelse hos tospråklige: En sammenfatning av empiriske studier. *Norsk pedagogisk tidsskrift*, *93*(04), 264-279.

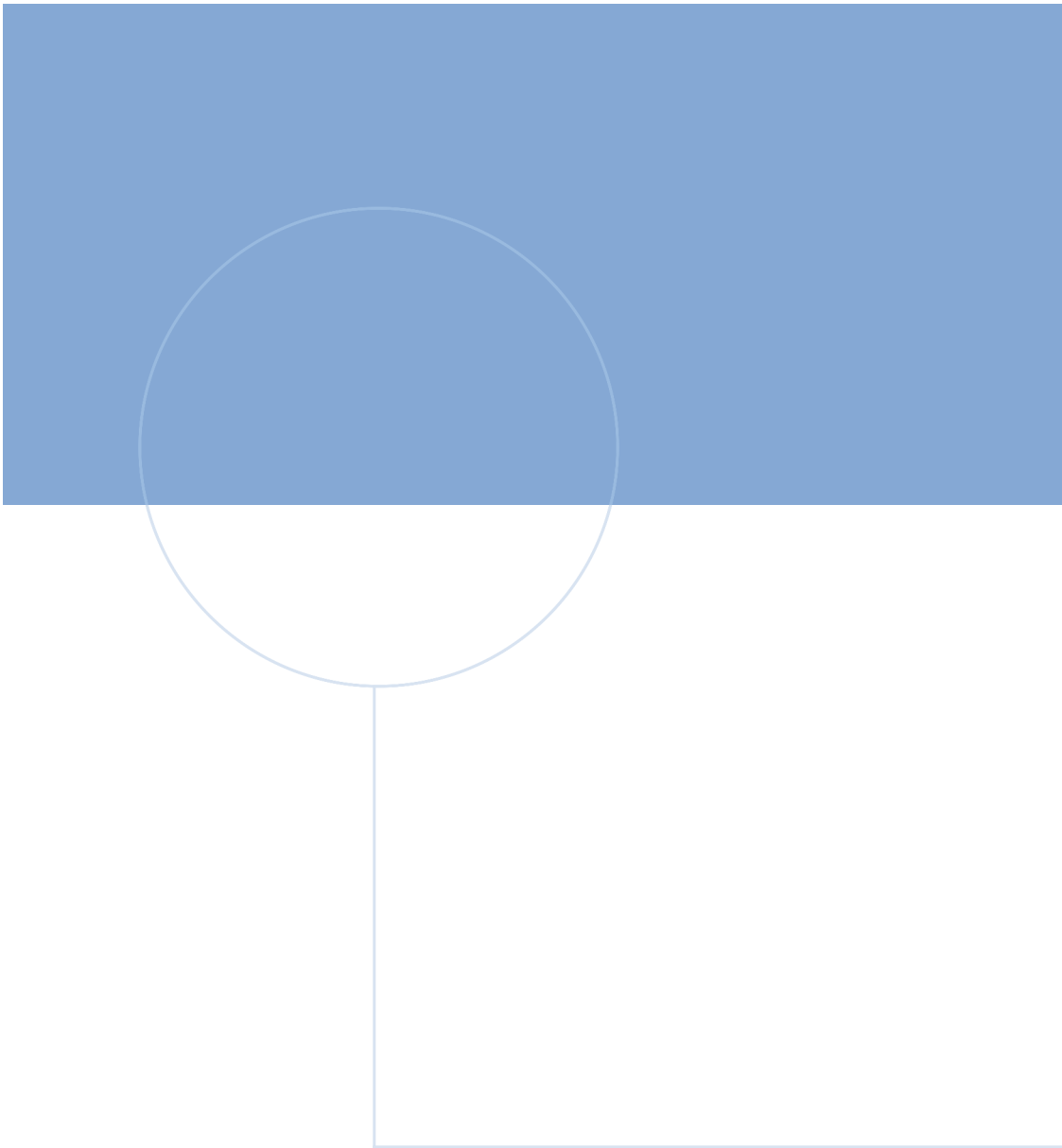
SSB. (2003). Inntekt og forbruk.

SSB. (2020). Registrerte arbeidsledige blant innvandrere.

Steinkellner, A. (2017). Hvordan går det med innvandrere og deres barn i skolen. *SSB. Artikkelserie: Innvandrere i Norge*.

Studenmund, A. H. (2017). *A Practical guide to using econometrics* (7th edition, global edition. utg.). Upper Saddle River.

Thomas, R. L. (2005). *Using statistics in economics*: McGraw-Hill.



NTNU

Kunnskap for en bedre verden