

Magnus Holten Skånlund

SØK2901 Bacheloroppgave i samfunnsøkonomi

Analyse av hvordan elevprestasjoner i land med ulikt velstandsnivå avhenger av kjønn, inntekt og klassestørrelse

Bacheloroppgave i MSØK/5

Veileder: Bjarne Strøm

Mai 2020

SØK2901 Bacheloroppgave i samfunnsøkonomi

Analyse av hvordan elevprestasjoner i land med ulikt velstandsnivå
avhenger av kjønn, inntekt og klassestørrelse

Magnus Holten Skånlund

Dato: 15.05.2020

Sammendrag

Elevprestasjoner i skolen kan være et viktig mål på fremtidig vekst i et land. Denne oppgaven ser på hvordan leseferdigheter hos 4. klasse-elever påvirkes av kjønn, klassestørrelse og inntekt. Gjennom analysen ser jeg at leseferdighetene variere mellom kjønnene og at denne forskjellen er uavhengig av BNP-nivået i landet. Videre ser jeg også at leseferdighetene varierer med klassestørrelse og familieinntekt, men for disse faktorene så vil effekten også påvirkes av landets BNP-nivå. I denne oppgaven gjør jeg flere typer tester for å underbygge konklusjonene nevnt over.

Innhold

SAMMENDRAG	2
INNHold	3
TABELLER OG FIGURER	4
1 INTRODUKSJON	5
1.1 MOTIVASJON.....	5
1.2 PROBLEMSTILLING	5
2 TEORETISK RAMMEVERK OG TIDLIGERE LITTERATUR	6
2.1 INNLEDNING.....	6
2.2 TEORETISK RAMMEVERK	6
2.3 TIDLIGERE LITTERATUR	7
3 DATAMATERIALET	8
3.1 INNLEDNING.....	8
3.2 OM DATAMATERIALET	8
3.2.1 Definisjon av avhengig variabel:.....	8
3.2.2 Definisjon av interessevariabler:	8
3.2.3 Definisjon av kontrollvariabler:.....	9
3.2.4 Definisjon av stokastisk restledd:	9
3.3 DESKRIPTIV STATISTIKK FOR AVHENGIG VARIABEL.....	9
3.4 DESKRIPTIV STATISTIKK FOR INTERESSEVARIABLE.....	11
3.5 DESKRIPTIV STATISTIKK FOR KONTROLLVARIABLE	13
3.6 KORRELASJONSMATRISSE FOR DATAMATERIALET.....	14
3.7 OPPSUMMERING	14
4 EMPIRISK STRATEGI	15
4.1 INNLEDNING.....	15
4.2 EMPIRISK STRATEGI	15
4.2.1 Valg av funksjonsform	15
4.2.2 Estimeringsmetode	16
4.3 OPPSUMMERING	16
5 EMPIRISKE RESULTATER	17
5.1 INNLEDNING.....	17
5.2 RESULTATER.....	17
5.2.1 Oversikt over modellvarianter som estimeres:	17
5.2.2 Estimering av modell (2)-(5)	18
5.2.3 Påvirkning av interessevariabler på leseferdigheter	19
5.2.4 Resultater for korrelasjon mellom BNP og interessevariablene.....	20
5.3 OPPSUMMERING	21
6 OPPSUMMERING OG KONKLUSJONER	22
6.1 BEGRENSNINGER OG MULIGE UTVIDELSER.....	22
REFERANSER:	23
APPENDIKS	24

Tabeller og figurer

- Tabell 1: Definisjon av kategorivariabel, *income*..... 8
 - Tabell 2: Definisjon av kontrollvariabel, "*par_emp*" 9
 - Tabell 3: Deskriptiv statistikk for den avhengige variabelen *read* 9
 - Tabell 4: Deskriptiv statistikk av de 5 landene som viser målinger av testscoren. 10
 - Tabell 5: Deskriptiv statistikk for interessevariablene..... 11
 - Tabell 6: Deskriptiv statistikk for testscore gitt elevens kjønn..... 13
 - Tabell 7: Deskriptiv statistikk for kontrollvariablene 13
 - Tabell 8: Korrelasjonsmatrise for utvalgte variabler 14
 - Tabell 9: Regresjonsresultater for modell (2)-(5) 18
 - Tabell 10: Resultat av hypotesetest for interessevariablene. 19
 - Tabell 11: T-test av interaksjonsledd..... 21
 - Tabell 12: Regresjon av modell (3) 24
 - Tabell 13: T-test av interessevariabler 24
 - Tabell 14: T-test av kvadratiske ledd..... 25
-
- Figur 1: Fordeling av avhengig variabel, *read*..... 9
 - Figur 2: Histogram som viser fordelingen av testscoren i de fem landene..... 10
 - Figur 3: Kryssplott av testscore mot klassestørrelse og en enkel regresjonslinje..... 11
 - Figur 4: Kryssplott av testscore mot velferdsnivå og en enkel regresjonslinje 11
 - Figur 5: Kryssplott av testscore mot inntekt i husstanden og en enkel regresjonslinje 12
 - Figur 6: Histogram som viser fordelingen i testscore for gutter 12
 - Figur 7: Histogram som viser fordelingen i testscore for jenter 12
 - Figur 8: BNP per capita i tusen USD..... 24
 - Figur 9: Kryssplott av *bnp* mot *clsiz* og en enkel regresjonslinje 24

1 Introduksjon

1.1 Motivasjon

Skolesystemet er viktig for et lands utvikling og i de senere årene har det blitt aksept for bruk av økonomiske modeller for å vurdere effektiviteten i skoler. En modell kan være å analysere en skoleproduktfunksjon der man ser på hvordan elevprestasjoner påvirkes av en rekke faktorer. Jeg vil gå nærmere inn på hvordan en skoleproduktfunksjon fungerer i kap. 2. Denne oppgaven vil bruke en slik økonometrisk modell for å se på forskjellen i elevprestasjoner mellom land med ulikt velstandsnivå, målt i BNP. Den vil også se på om det foreligger forskjeller i hvor mye faktorer som kjønn, inntekt og klassestørrelse påvirker prestasjonene i de 5 landene.

1.2 Problemstilling

Hvordan påvirkes elevprestasjoner av kjønn, inntekt og klassestørrelse, og varierer dette med velstandsnivået i landet?

Utvalgte land med forskjellig velstandsnivå (*se appendiks, figur 8 for fordeling*), som blir analysert i oppgaven:

- Norge
- England
- Tsjekkia
- Bulgaria
- Belize

2 Teoretisk rammeverk og tidligere litteratur

2.1 Innledning

Dette kapitlet gir en generell beskrivelse av skoleproduktfunksjonen, et overblikk over dataene oppgaven bygger på, samt tidligere relevant litteratur og resultater.

2.2 Teoretisk rammeverk

Når et land skal se på hvordan man kan øke prestasjonene i skolen har det blitt vanlig å bruke økonomiske modeller for nyttemaksimering. Gjennom modeller som skoleproduktfunksjonen kan man få oversikt over hvilke faktorer som påvirker elevenes resultater, og hvor mye. Politikere vil da ha et bedre grunnlag for å innføre skolereformer som gir elevene et bedre utbytte av skolegangen.

Denne oppgaven bruker skoleproduktfunksjonen som utgangspunkt og grunnmodell for analysen. Det betyr at jeg ser resultatene i lys av en rekke faktorer. En enkel skoleproduktfunksjon kan beskrives som følgende:

$$(1) T = f(S, F, P)$$

T = Testscore

F = Familie-/elevkarakteristika (inntekt, kjønn osv.)

P = Medelevkarakteristika («peer group» effekter)

S = Skolefaktorer (lærerkarakteristika, klassestørrelse osv.)

Modellen ser kanskje simpel ut, men som jeg skal se videre kan en slik modell være med på å fange opp mange forskjellige påvirkningsfaktorer på testscoren.

Når jeg senere i oppgaven skal se på effektene av forskjellige variabler på testscoren kommer jeg til å bruke T-test for å verifisere resultatene våre. En T-test er en form for hypotesetest som brukes når man kun ser på en variabel om gangen. Dersom man skal ha med flere variabler i testen må man bruke en F-test. Slike tester vil bli forklart senere i oppgaven.

2.3 Tidligere litteratur

Hans Bonesrønning¹ gir en innføring i hvordan man kan bruke økonomiske modeller til å øke prestasjonene i skolen. Han tar utgangspunkt i en produktfunksjon for utdanning med en rekke variabler som påvirker prestasjonen til elever. Dette er variabler som tar høyde for familiekarakteristika, «peer group» effekter (påvirkning fra medelever) og skolefaktorer, samt et restledd som fanger opp alle andre faktorer utenfor modellen. En slik skoleproduktfunksjon, finner Bonesrønning, er hensiktsmessig for å se på hvordan man kan maksimere «produksjonen» av resultater hos elevene.

Eric A. Hanushek² går mer inn på styrker og svakheter ved skoleproduktfunksjoner. Han påpeker at det er mange som måler ferdigheter etter hvor mange års skolegang en elev har. Altså at en elev med 13 års skolegang er smartere enn en med 11 års skolegang. Hanushek mener dette er en for enkel og lite treffsikker måte å måle ferdigheter på. Han mener at det i stedet bør gjøres nøye uttenkte tester, som blir gjort likest mulig på tvers av land. Man kan da få grunnlag for å vurdere om antall års skolegang påvirker prestasjonen på testene. Hanushek konkluderer til slutt med at det er vanskelig å si noe sikkert om hvilke faktorer som påvirker prestasjonen mest, ut ifra empirien. Problemet er at det blir vanskelig å måle prestasjonene av et representativt utvalg når antallet faktorer er så stort. Det eneste man kan si med moderat sikkerhet er at det ikke er mengden ressurser som blir brukt som er viktig, men heller hvordan disse ressursene blir brukt.

Rapporten fra Stoltenbergutvalget³ er skrevet av et ekspertutvalg som skulle se på forskjeller i skoleprestasjoner mellom kjønnene, og hvorfor de oppstår. I forbindelse med denne oppgaven er det relevant å presentere de resultatene som sier noe om forskjellene i elevprestasjon mellom kjønn. Det de fant var at det er en betydelig forskjell mellom gutter og jenter når det kommer til leseferdigheter. De hadde da basert seg på PIRLS-undersøkelsen og konkluderte med at jenter generelt gjør det bedre enn gutter, i hvert fall på et 4.klassenivå.

¹ Bonesrønning 2004: s. 14-23

² Hanushek 2020: s. 161-170

³ Stoltenberg et al. 2019: s. 14-15

3 Datamaterialet

3.1 Innledning

I dette kapitlet presenteres datamaterialet. Jeg vil gi litt informasjon om studiet som ble gjort rundt materialet opprinnelig, før jeg går videre til å se på de viktigste variablene i modellen vår.

3.2 Om datamaterialet

Datamaterialet som blir brukt i denne oppgaven er hentet fra en internasjonal undersøkelse av leseferdigheter hos 4. klasseelever kalt Progress in International Reading Literacy Study (PIRLS)⁴. PIRLS ble gjennomført i 2001 av International association for the evaluation of educational achievement, IEA. Dataene som blir brukt i denne oppgaven er tilrettelagte data fra grunnlagsmaterialet i PIRLS, 2001. Det inneholder data for 35 forskjellige land og omlag 150 000 elever.

De aktuelle datafilene som brukt i denne oppgaven er:

- data2901-Bulgaria.dta
- data2901-Norway.dta
- data2901-England.dta
- data2901-CzechRepublic.dta
- data2901-Belize

3.2.1 Definisjon av avhengig variabel:

Read = testscore på test i leseferdigheter

3.2.2 Definisjon av interessevariabler:

Girl = dummyvariabel som forteller kjønnet på eleven

$girl = 1 \rightarrow$ jente

$girl = 0 \rightarrow$ gutt

Income = kategorivariabel som viser årlig inntekt i husstanden, gitt i USD

LESS THAN \$20,000	1
\$20,000 - \$29,999	2
\$30,000 - \$39,999	3
\$40,000 - \$49,999	4
\$50,000 - \$59,000	5
\$60,000 OR MORE	6

Tabell 1: Definisjon av kategorivariabel, *income*

Clsize = klassestørrelse

BNP = variabel som viser BNP-nivået til hvert av landene i datasettet

⁴ Senter for leseforskning: side 5

3.2.3 Definisjon av kontrollvariabler:

Par_emp = kategori som viser hvor mye foreldrene jobber

Both work fulltime	1
One (not both) work fulltime	2
Both working less than fulltime	3
Not working	4

Tabell 2: Definisjon av kontrollvariabel, "*par_emp*"

Jeg gjør om denne kategorivariabelen til en dummy for å lettere kunne se på effekten av foreldrenes grad av sysselsetting. Jeg får da variabelen *par_emp_high*. Dersom den er lik 1 betyr det at foreldrene jobber lite (nivå 3-4 i tabell 2) og hvis *par_emp_high*=0 jobber foreldrene mye (nivå 1-2 i tabell 2).

Teacher_exp = viser antall år som lærer

3.2.4 Definisjon av stokastisk restledd:

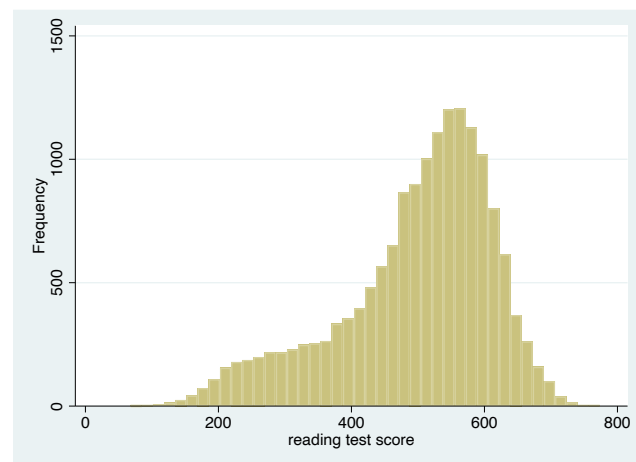
ϵ = stokastisk restledd, fanger opp variasjoner i modellen som de andre variablene ikke inkluderer.

3.3 Deskriptiv statistikk for avhengig variabel

Den deskriptive statistikken nedenfor viser den avhengige variabelen, først for alle landene samlet og så for hvert enkelt land.

<i>Read</i>	
Gjennomsnitt	496,5
Standardavvik	116,2
Max	773,4
Min	67,9
Observasjoner	16 006

Tabell 3: Deskriptiv statistikk for den avhengige variabelen *read*



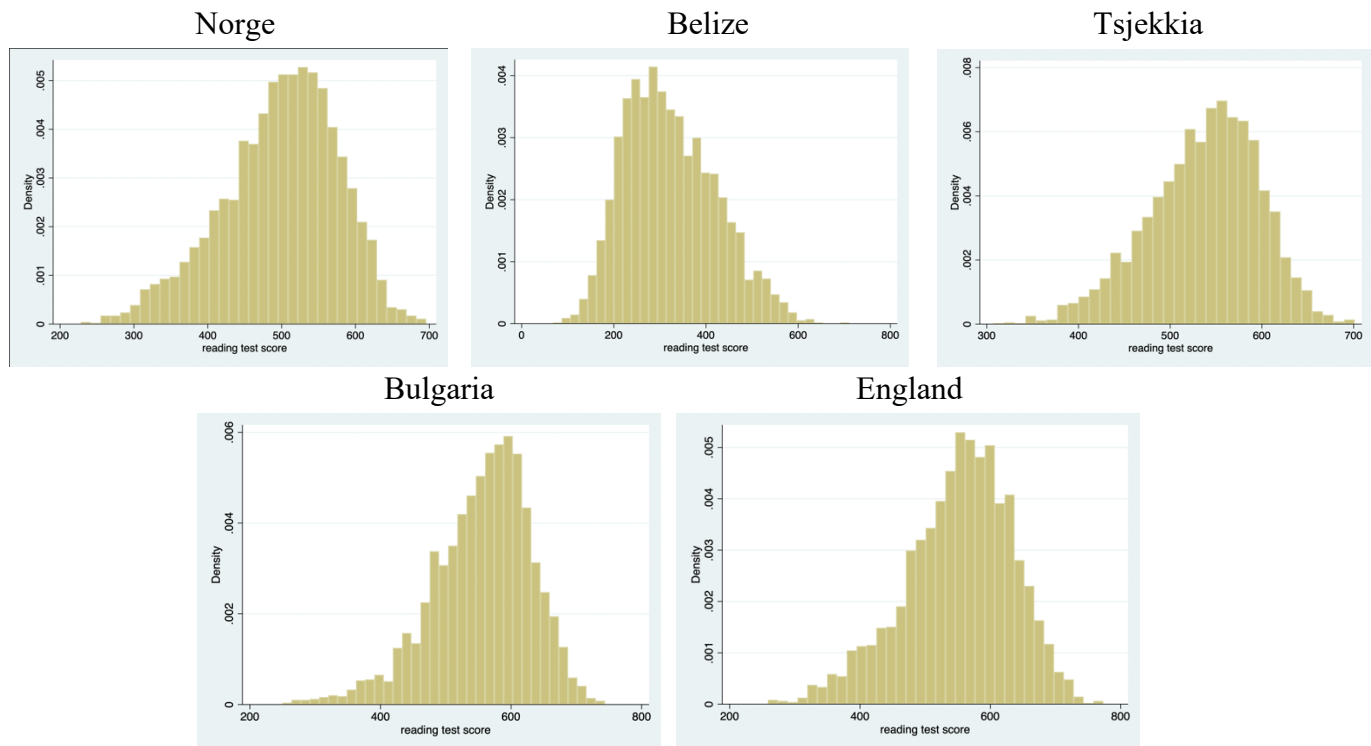
Figur 1: Fordeling av avhengig variabel, *read*.

I tabell 4 og figur 2 ser jeg på ulike statistikker for den avhengige variabelen i de fem landene. Her ser jeg at det ikke nødvendigvis er noen korrelasjon mellom *BNP* og resultatene på

lesetesten. Jeg ser at Bulgaria som har den høyeste testscoren har nest lavest nivå på *BNP*. Norge som har det høyeste velstandsnivået har nest lavest testscore. Jeg skal se nærmere på denne korrelasjonen i kapittel 5.

	Norge	Belize	Tsjekkia	Bulgaria	England
Gjennomsnitt	498,3	323,2	536,7	555	551,5
Standardavvik	78,4	99,1	60,7	76,1	82,7
Max	695,9	710,8	701,2	743,7	773,4
Min	228,1	67,9	307,3	249,7	258,6
Observasjoner	3 459	2 909	3 022	3 460	3 156
BNP	78,33	4,98	24,57	10,13	40,39

Tabell 4: Deskriptiv statistikk av de 5 landene som viser målinger av testscoren.



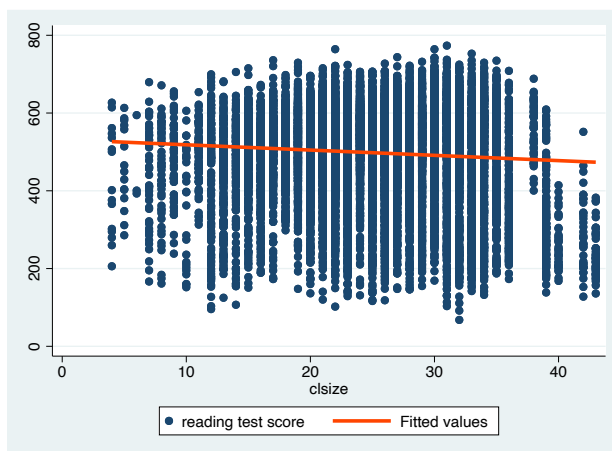
Figur 2: Histogram som viser fordelingen av testscoren i de fem landene

3.4 Deskriptiv statistikk for interessevariable

Her presenteres det deskriptiv statistikk for de fire interessevariablene. Jeg skal også se på kryssploott med regresjonslinje på variablene *income*, *clsiz* og *BNP* mot *read* (se figur 3-5).

	<i>girl</i>	<i>income</i>	<i>clsiz</i>	<i>BNP</i>
Gjennomsnitt	0,5	2,4	24,7	32,6
Standardavvik	0,5	1,7	6,2	26,9
Max	1	6	43	78,3
Min	0	1	4	4,98
Observasjoner	15 777	11 029	15 288	16 006

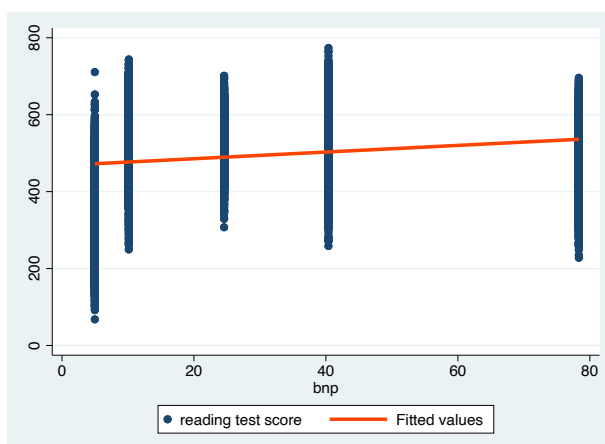
Tabell 5: Deskriptiv statistikk for interessevariablene



Figur 3: Kryssploott av testscore mot klassestørrelse og en enkel regresjonslinje

Kommentar:

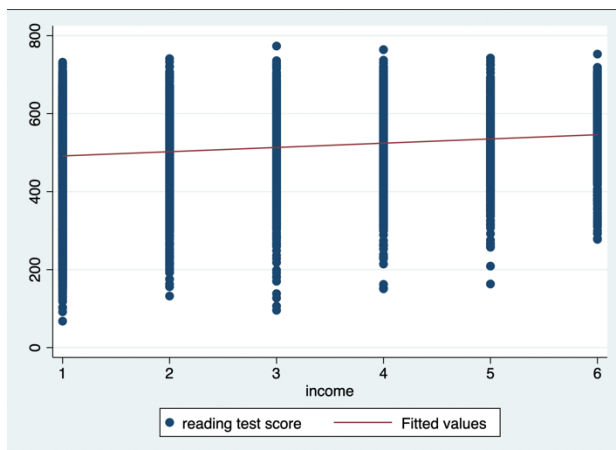
Figur 3 illustrerer at klassestørrelse og testscore er negativt korrelert, med partiell korrelasjonskoeffisient på $-0,073$. Det vil si at dersom man reduserer klassestørrelsen med 1 elev så vil testscoren forbedres $0,073$.



Figur 4: Kryssploott av testscore mot velferdsnivå og en enkel regresjonslinje

Kommentar:

Figur 4 illustrerer at *BNP* og testscore er positivt korrelert, med en partiell korrelasjonskoeffisient på $0,067$ (jfr. tabell 8). Det vil si at dersom *BNP* øker med 1 enhet (1000USD) vil testscoren øke med $0,067$.



Figur 5: Kryssplott av testscore mot inntekt i husstanden og en enkel regresjonslinje

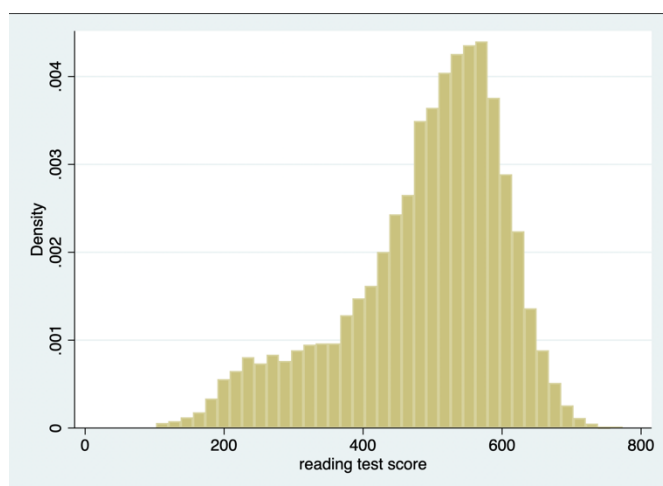
Kommentar:

Figur 5 illustrerer at inntekten i husstanden og testscore er positivt korrelert, med en partiell korrelasjonskoeffisient på 0,168. Det vil si at dersom *income* øker med 1 enhet (vil si at man går fra en inntektsklasse over i den neste) vil testscoren øke med 0,168.

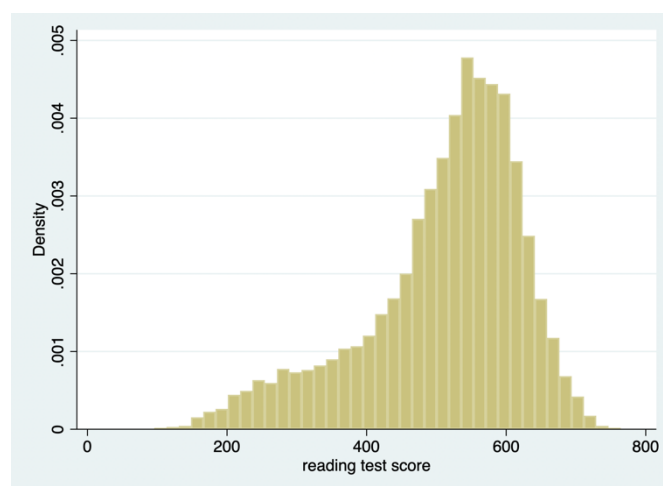
Resultatene fra figur 3-5 kan ikke regnes som en kausal sammenheng fordi det er sp mange andre variabler som også kan påvirke *read*.

Kryssplottene kunne vist en tydeligere sammenheng dersom variablene var kontinuerlige. Nå som de er kategoriske ser jeg at de ikke har en jevn spredning av punkter. Jeg velger likevel å vise plottene fordi de visuelt viser sammenhenger og trender, selv om trendene er tilsynelatende svake.

Jeg går nå videre for å se på statistikk for dummy-variabelen **girl**. Jeg bruker *girl*=0 for gutter og *girl*=1 for jenter, og setter dem inn i et histogram for å se på fordelingen av de to.



Figur 6: Histogram som viser fordelingen i testscore for gutter



Figur 7: Histogram som viser fordelingen i testscore for jenter

Tabellen under viser mer deskriptiv forskjellen mellom leseferdighetene til gutter og jenter: Her foreligger det en tilsynelatende betydelig forskjell i resultatene for gutter og jenter, jeg skal i senere kapittel se på om denne differansen er signifikant eller ikke. Det er også interessant å se på om den samme forskjellen mellom gutter og jenter varierer med BNP-nivå.

	Jenter	Gutter
Gjennomsnitt	509,3	487,6
Standardavvik	113,3	115,9
Max	764,2	773,4
Min	95,7	102,1
Observasjoner	7 837	7 940

Tabell 6: Deskriptiv statistikk for testscore gitt elevens kjønn

3.5 Deskriptiv statistikk for kontrollvariable

Jeg har allerede presentert kontrollvariablene som skal brukes i estimeringen. Jeg vil her presentere deskriptiv statistikk for disse variablene:

	<i>Par_emp_high = 0</i>	<i>Par_emp_high = 1</i>	<i>Teacher_exp</i>
Gjennomsnitt	465,8	519,5	15,6
Standardavvik	126,9	103,4	10,8
Max	729,5	773,4	53
Min	67,9	103,4	1
Observasjoner	1 540	10 131	15 076

Tabell 7: Deskriptiv statistikk for kontrollvariablene

Forklaring av tabell 7:

Jeg ser av tabellen at gjennomsnittlig testscore for de barna hvor foreldrene har høy grad av sysselsetting (nivå 1-2 i tabell 2) er gjennomsnittlig testscore på 519,5. Jeg ser også at de elevene med foreldre som har lav sysselsetting (nivå 3-4 i tabell 2) har et snitt på 465,8. En svakhet ved datasettet er at gruppen elever som har høy grad av sysselsetting er stor, mens gruppen elever med foreldre med lav grad av sysselsetting er relativt liten. Med over 1500 i den minste gruppen antar jeg imidlertid at også denne gruppen har et representativt utvalg. Gjennomsnittlig lærer-erfaring(*teacher_exp*) ser jeg av tabell 7 er på 15,6 år.

3.6 Korrelasjonsmatrise for datamaterialet

Under presenterer jeg en matrise med korrelasjonskoeffisientene til de utvalgte variablene:

	read	girl	income	clsize	bnp	par_em~h	teache~p
read	1.0000						
girl	0.0874	1.0000					
income	0.1349	0.0060	1.0000				
clsize	-0.0441	0.0067	-0.0141	1.0000			
bnp	0.0663	-0.0159	0.6752	-0.2058	1.0000		
par_emp_high	0.1606	-0.0086	0.2338	0.0503	0.2093	1.0000	
teacher_exp	0.1120	0.0133	-0.0180	-0.0706	0.0156	0.0251	1.0000

Tabell 8: Korrelasjonsmatrise for utvalgte variabler

Kommentar:

Jeg ser fra tabell 8 at flere av variablene har en svak korrelasjon med den avhengige variabelen *read* og med interessevariablene *girl*, *income*, *clsize* og *BNP*.

En korrelasjonsmatrise ser på om variablene er lineært uavhengige av hverandre. Verdiene i matrisen viser korrelasjonen mellom to variabler, gitt ved R . Dersom $R=1$ eller $R=-1$ er variablene perfekt positivt/negativt korrelert. Et problem ved en slik korrelasjonsmatrise er at den ikke forklarer hvor mye variabelen verdi kan forklares av den andres variasjon. For å se på forklaringskraften mellom variablene må jeg se på R^2 .

Jeg ser av tabell 8 at det ikke er spesielt høy korrelasjon mellom noen av variablene. Det eneste stedet med relativt høy korrelasjon er mellom *BNP* og *income*. Dette kan føre til et multikollinearitetsproblem, som vil si at variablene påvirker hverandre såpass mye at de til en viss grad blir overflødige.

En utfordring som kan oppstå er et utelatt variabelproblem. Problemet oppstår når man utelater en variabel som viser seg å ha betydning for utfallet av analysen. Dette kan vise seg gjennom en skjevhet i resultatet. Kontroll- eller interessevariablene kan ende opp med å få større eller mindre forklaringskraft enn den i realiteten skulle hatt. Noe som kan føre til en feilestimering av modellen.

3.7 Oppsummering

Jeg har i dette kapitlet presentert datamaterialet fra rapporten PIRLS. Jeg har sett på statistikk for de forskjellige variablene, samt hvordan de samhandler med hverandre. I neste kapittel tar jeg for meg strategien for gjennomføring av den empiriske analysen.

4 Empirisk strategi

4.1 Innledning

Dette kapitlet viser min plan for analysen av datamaterialet. Jeg presenterer variablene og forklarer metodene jeg bruker.

4.2 Empirisk strategi

Min oppgave har en naturlig to-delning. Jeg vil derfor først gjøre tre estimeringer med fokus på interessevariablene for kjønn, inntekt og klassestørrelse. Deretter vil jeg se på om estimatene varierer mellom landene i utvalget.

Det teoretiske rammeverket bunner kort fortalt ut i skoleproduktfunksjonen (1), og spesielt hvordan flere variabler enn kun interessevariabelen må vurderes for å få et helhetlig bilde av hvordan ferdighetene påvirkes. Jeg vil derfor også trekke inn andre forhold enn kun kjønn, inntekt og klassestørrelse i en multippel regresjonsmodell for å estimere effektene på leseferdighetene.

4.2.1 Valg av funksjonsform

Jeg starter ut med en lineær tilnærming av (1) før jeg senere skal se på ikke-lineære videreføring av modellen:

$$(2) \text{ read} = \beta_1 + \beta_2 \text{ girl} + \beta_3 \text{ income} + \beta_4 \text{ clsiz} + \beta X + \varepsilon$$

Der *read* er målet på leseferdigheter, *girl* er mål på kjønn, *income* er mål på inntekt i husstand(USD), *clsiz* er mål på klassestørrelse og *X* er en samling av kontrollvariable med koeffisientvektor β . ε er et stokastisk restledd som fanger opp alle de andre variablene som ikke er spesifisert i modellen, men som påvirker resultatet. $\beta_2, \beta_3, \beta_4$ måler marginaleffekten av parameteren gitt at de andre variablene er konstante. Dersom man kvadrerer et slikt ledd vil man se hvordan marginaleffekten endrer seg.

$$(3) \text{ read} = \beta_1 + \beta_2 \text{ girl} + \beta_3 \text{ income} + \beta_4 \text{ clsiz} + \beta_5 \text{ par_emp_high} + \beta_6 \text{ teacher_exp} + \varepsilon$$

Jeg trekker deretter inn et par kontrollvariable til for å få en mer korrekt måling av koeffisientvektorene. Grunnen til at jeg tar med kontrollvariable er for å gi modellen høyere forklaringskraft. Det er ganske intuitivt at dersom man har med få variabler blir det vanskelig å si mye om effektene på leseferdigheter. Når man derimot tar med flere variabler er det enklere å sammenlikne effekten av disse på hverandre og på leseferdigheter og dermed få et mest mulig korrekt resultat.

$$(4) \text{ read} = \beta_1 + \beta_2 \text{ girl} + \beta_3 \text{ income} + \beta_4 \text{ clsiz} + \beta_5 \text{ par_emp_high} + \beta_6 \text{ teacher_exp} + \gamma_1 \text{ income_sq} + \gamma_2 \text{ clsiz_sq} + \varepsilon$$

Modell 4 ser på hvordan effekten av *income* og *clsiz* på *read* utvikler seg. Jeg har da generert kvadratiske ledd av de to variablene. Koeffisientvektoren til γ_1 og γ_2 vil fortelle oss om effekten av variabelen er avtagende, stigende eller konstant. Jeg har ikke laget et kvadratisk

ledd av variabelen *girl* fordi det ikke er hensiktsmessig å se på et kvadratisk ledd av en dummyvariabel.

$$(5) \text{ read} = \beta_1 + \beta_2 \text{ girl} + \beta_3 \text{ income} + \beta_4 \text{ clsize} + \beta_5 \text{ par_emp_high} + \beta_6 \text{ teacher_exp} + \beta_6 \text{ bnp} + \delta_1 \text{ girl_bnp} + \delta_2 \text{ income_bnp} + \delta_3 \text{ clsize_bnp} + \varepsilon$$

I modell 5 har jeg tatt med interaksjonsledd mellom de tre interessevariablene og *BNP* for å kunne se på koeffisientvektorene δ_1 , δ_2 og δ_3 . Interaksjonsleddene vil si noe om BNP-nivået i et land har noen innvirkning på effekten av interessevariablene på leseferdigheter. Jeg vil da se om det foreligger ikke-linearitet i effekten av β -koeffisientene fra modell 3.

4.2.2 Estimeringsmetode

Det viktigste blir å estimere parameterne β_2 , β_3 og β_4 i (2) og foreta hypotesetester rundt disse. Jeg vil da bruke minste kvadraters metode (OLS) som estimeringsmetode. Forutsetningene for å kunne bruke OLS er at regresjonskurven er lineær, normalfordelt og homoskedastisk (konstant varians) restledd som er uavhengig mellom de forskjellige observasjonene. Restleddet må være normalfordelt for å kunne bruke T- og F-tester til å hypoteseteste parameterne i modellen. Den siste viktige forutsetningen er at ε er ukorrelert med forklaringsvariablene i modellen. Under disse forutsetningene vil OLS-estimatorene til β_2 , β_3 og β_4 være forventningsrette og variansminimale. Det vil si at estimatoren ikke systematisk avviker fra den parameteren jeg ønsker å estimere og at estimatene er effisiente.

Regresjonen vil bli gjort i STATA og dataene blir presentert i en tabell som gir grunnlaget for å tolke de forskjellige koeffisientvektorene.

4.3 Oppsummering

Jeg har nå vist hvordan jeg skal gå fram for å finne effektene av interessevariablene på leseferdigheter. Under estimeringen vil det være viktig å se på tester rundt parameterne β_2 , β_3 og β_4 i ligning (2). Jeg vil så gå videre og se på om estimatene innad i de utvalgte landene varierer betydelig fra disse estimatene, dette vil gjøres gjennom modell (4) der jeg har inkludert interaksjonsledd mellom interessevariablene og *BNP*.

5 Empiriske resultater

5.1 Innledning

Jeg vil i dette kapitlet presentere resultater fra de forskjellige delene av analysen. Jeg starter med å presentere de modellene som skal estimeres. Jeg vil videre analysere modell (2) og (3) for å vise hvordan interessevariablene generelt (de fem landene kombinert) påvirker leseferdighetene. Som en videreføring av dette vil jeg i (4) inkludere kvadratiske ledd av *income* og *clsize*, slik at jeg kan vurdere om effekten av variablene endres over tid eller om de er lineære. Jeg betrakter altså ikke-linearitet i effekten av de to variablene. Videre vil jeg se på om det foreligger en trend i påvirkningskraften til interessevariablene på prestasjonen, ut fra BNP-nivået på landene. Dette gjør jeg ved å betrakte modell (5).

5.2 Resultater

5.2.1 Oversikt over modellvarianter som estimeres:

$$(2) \text{ read} = \beta_1 + \beta_2 \text{ girl} + \beta_3 \text{ income} + \beta_4 \text{ clsize} + \varepsilon$$

Lineær grunnmodell

$$(3) \text{ read} = \beta_1 + \beta_2 \text{ girl} + \beta_3 \text{ income} + \beta_4 \text{ clsize} + \beta_5 \text{ par_emp_high} + \beta_6 \text{ teacher_exp} + \varepsilon$$

Lineær grunnmodell med kontrollvariabler

$$(4) \text{ read} = \beta_1 + \beta_2 \text{ girl} + \beta_3 \text{ income} + \beta_4 \text{ clsize} + \beta_5 \text{ par_emp_high} + \beta_6 \text{ teacher_exp} + \gamma_2 \text{ income_sq} + \gamma_3 \text{ clsize_sq} + \varepsilon$$

Ser på ikke-linearitet ved å ta med kvadratisk ledd av interessevariablene *clsize* og *income*. Forteller om endring i effekten til variabelen.

$$(5) \text{ read} = \beta_1 + \beta_2 \text{ girl} + \beta_3 \text{ income} + \beta_4 \text{ clsize} + \beta_5 \text{ par_emp_high} + \beta_6 \text{ teacher_exp} + \beta_6 \text{ bnp} + \delta_1 \text{ girl_bnp} + \delta_2 \text{ income_bnp} + \delta_3 \text{ clsize_bnp} + \varepsilon$$

Modell som ser på om det foreligger ikke-linearitet i effekten mellom *BNP* og de tre interessevariablene. Inkluderer interaksjonsledd for interessevariablene i kombinasjon med *BNP* per capita.

Disse modellene er allerede beskrevet i detalj i kapittel 4 så jeg går derfor videre for å se på regresjonen av disse modellene.

5.2.2 Estimering av modell (2)-(5)

	2.	3.	4.	5.
Variables	read	read	read	read
girl	18.72	18.49	18.96	19,05
	(2.106)	(2.117)	(2.052)	(3.222)
clsize	-1.128	-0.748	21.91	-6.223
	(0.172)	(0.176)	(0.924)	(0.264)
income	9.700	6.326	10,32	22,06
	(0.610)	(0.624)	(2.909)	(1.402)
par_emp_high		43.73	38.03	47.48
		(3.282)	(3.216)	(3.165)
teacher_exp		1.069	0.959	1.010
		(0.100)	(0.0973)	(0.0963)
income_sq			-0.601	
			(0.440)	
clsize_sq			-0.470	
			(0.0189)	
bnp				-2.764
				(0.162)
girl_bnp				-0.0190
				(0.0702)
clsize_bnp				0.153
				(0.00618)
income_bnp				-0.315
				(0.0251)
_cons	504.2	452.6	199.0	554.7
	(4.658)	(5.562)	(11.81)	(7.389)
N	10525	9541	9541	9541
R²	0.035	0.057	0.115	0.133

Tabell 9: Regresjonsresultater for modell (2)-(5)

Forklaring tabell 9: viser endring i *read* dersom det er en enhets økning i venstresidevariablene.

N = Antall observasjoner.

R^2 = Determinasjonskoeffisienten (hvor stor grad av variasjonen i *read* som forklares av modellen)

Tallene i parentes er estimerte standardavvik for den variabelen. De er gitt ved å ta kvadratroten av $\sigma_{b_2}^2 = \frac{\sigma}{\sum x_2^2(1-r^2)}$ (formel for å finne variansen til β_2). Der σ^2 er variansen til restleddet. X er variabelen i fokus, r er korrelasjonen mellom variabelen og den forrige variabelen.

5.2.3 Påvirkning av interessevariabler på leseferdigheter

Jeg skal videre foreta tre hypotesetester, en for hver av de tre interessevariablene. Dette gjør jeg ved å se på P-verdien til de tre variablene hentet fra regresjonen av den utvidede grunnmodellen (3). P-verdien forteller oss om sannsynligheten for at det er feil å forkaste nullhypotesen. Dersom p-verdien er lavere enn det valgte signifikansnivået forkastes hypotesen. Nullhypotesen til de tre variablene vil være « β -verdien til variabelen» = 0, og alternativhypotesen; « β -verdien til variabelen» \neq 0. De tre hypotesetestene vil da bli:

$H_0: \beta_2=0, H_A: \beta_2\neq 0$ $H_0: \beta_3=0, H_A: \beta_3\neq 0$ $H_0: \beta_4=0, H_A: \beta_4\neq 0$

Som vil si at dersom jeg forkaster nullhypotesen (at β -verdien til variabelen $\neq 0$) er det en statistisk signifikant effekt av interessevariabelen på *read*.. Et signifikansnivå er et uttrykk for om det er sannsynlig at resultatet er gitt av en tilfeldighet. Jo lavere signifikansnivå jo bedre. Det er vanlig å bruke et 95% konfidensintervall og dermed et signifikansnivå på 0,05, eller 5%. Jeg vet derfor at dersom en av interessevariablene har en p-verdi på under 0,05 forkastes den.

Dataene i tabellen under er hentet fra tabell 12 som ligger i appendiksen.

	<i>girl</i>	<i>income</i>	<i>clsiz</i>
P-verdi	0	0	0,001
Forkastes?	JA	JA	JA

Tabell 10: Resultat av hypotesetest for interessevariablene.

girl: Gjør jenter det bedre enn gutter?

Jeg kan allerede se fra tabell 6 at jenter har et høyere gjennomsnittlig resultat enn gutter. Og når jeg foretar en regresjon av modell (3) ser vi, jfr. tabell 9, at; det å være jente gir en økning i testresultatet på 18,35 i forhold til gutter.

Fra tabell 10 ser jeg at jeg kan forkaste nullhypotesen $H_0: \beta_2 = 0$. Det vil si at effekten av *girl* på *read* er signifikant på 0,1% signifikansnivå. Det er derfor god grunn for å si at jenter generelt gjør det bedre enn gutter på lesetesten.

income: *Gjør man det bedre på testen hvis man kommer fra en familie med høy inntekt?*

Fra tabell 8 ser jeg at det er en positiv korrelasjon mellom *income* og *read*. Og at resultatet på lesetesten øker med 5,8 (jfr. tabell 9) når man går fra en inntektsklasse til den over. Fra tabell 10 ser jeg at jeg kan forkaste nullhypotesen $H_0: \beta_3 = 0$. Det vil si at effekten av *income* på *read* er signifikant på alle signifikansnivå. Jeg kan derfor si at man generelt gjør det bedre på lesetesten dersom man kommer fra en familie med høy inntekt.

Jeg ser også at koeffisientvektoren til at det kvadratiske leddet *income_sq* i tabell 9, kolonne 3, er negativ. Det betyr at effekten av *income* på *read* er avtagende. Ser av tabell 14 i appendiks at denne effekten også er signifikant på alle signifikansnivå

clsiz: *Forbedres leseferdighetene dersom man reduserer klassestørrelsen?*

Jeg ser igjen fra tabell 8 at det er en negativ korrelasjon mellom *clsiz* og *read*, som vil si at dersom klassestørrelsen reduseres, forbedres resultatet på testen med. Det vil si at dersom man reduserer antallet elever i klassen med en, øker resultatet med 0,6 poeng (jfr. tabell 9).

Fra tabell 10 ser jeg at jeg kan forkaste nullhypotesen $H_0: \beta_4 = 0$. Det vil si at effekten av *clsiz* på *read* er signifikant på alle signifikansnivå. Det er altså grunnlag for å si at leseferdighetene forbedres dersom man reduserer klassestørrelsen.

Jeg ser også at koeffisientvektoren til at det kvadratiske leddet *clsiz_sq* i tabell 9, kolonne 3, er negativ. Det betyr at effekten av klassestørrelse på *read* er avtagende. Ser av tabell 14 i appendiks at denne effekten også er signifikant på alle signifikansnivå

5.2.4 Resultater for korrelasjon mellom *BNP* og interessevariablene

Er effekten av interessevariablene forskjellig ut ifra nivået på BNP i landet?

Jeg har tidligere i oppgaven (jfr. tabell 4 og figur 2) sett at det foreligger en forskjell i testscoren mellom landene med ulikt *BNP*-nivå. Men som jeg ser av tabell 4 så er det ingen åpenbar korrelasjon i effekten av *BNP* på testscoren. Det jeg ser på i dette delkapitlet er om *BNP* påvirker interessevariablene på en systematisk måte. Dermed gjør jeg litt som i forrige kapittel og ser på korrelasjonen til de tre interessevariablene og *BNP* hver for seg.

girl: *Foreligger det en korrelasjon mellom BNP og girl?*

Starter med å se på korrelasjonsmatrisen at det foreligger en initiell korrelasjon mellom *BNP* og *girl* på (fra tabell 8) $R = -0,0159$. Forklaringskraften mellom de to variablene blir da så lav at jeg kan si at det ikke foreligger en samvariasjon mellom de to variablene. Det vil altså si at effekten av kjønn på leseferdigheter ikke påvirkes av *BNP*-nivået i landet.

Jeg viser da til T-testen i tabell 11 som viser at sannsynligheten for at $H_0: girl_bnp = 0$ ikke kan forkastes er på cirka 80%. Jeg måtte dermed hatt et signifikansnivå på over 0,8 for å kunne forkaste H_0 . Det betyr at de to variablene ikke påvirker hverandre nevneverdig og jeg kan dermed si at hypotesen ikke kan forkastes. Det foreligger altså ingen korrelasjon mellom effekten av *girl* på *read* og *BNP*.

income: Foreligger det en korrelasjon mellom BNP og *income*?

For *income* ser jeg igjen først på tabell 8 at korrelasjonen mellom BNP og *income* er på $R=0,6752$. Får da $R^2=0,456$, som er relativt høyt. Jeg kan så se fra tabell 11 at jeg kan forkaste nullhypotesen om ikke-linearitet med 1% signifikansnivå. Det er derfor god grunn til å si at BNP og inntekten i husstanden korrelerer. Dette er jo også en rimelig konklusjon ettersom det er rimelig å anta at jo høyere BNP er i et land, jo høyere vil inntekten i husstandene være.

clsize: Foreligger det en korrelasjon mellom BNP og *clsize*? Korrelasjonen mellom *clsize* og *bnp* er ifølge tabell 8 på $R=-0,2058$, som gir $R^2=0,0424$. Det betyr at det er en viss negativ korrelasjon mellom den, men det er ikke mye av denne korrelasjonen som kommer av samvariasjonen mellom de to variablene. Fra tabell 11 ser jeg at jeg kan forkaste nullhypotesen om ikke-linearitet på et 1% signifikansnivå. Det foreligger altså en negativ korrelasjon mellom BNP og klassestørrelse, og som jeg ser av figur 9 i appendiksen virker det riktig. I datasettet foreligger det altså en trend der jo høyere nivå et land har på BNP dess mindre er klassestørrelsen.

5.3 Oppsummering

I dette kapitlet har jeg sett på resultatene for modellene presentert i oppgaven. Jeg har sett at leseferdighetene påvirkes av alle de tre interessevariablene. Effekten av variablene på leseferdigheter blir påvirket av BNP-nivået i landet for klassestørrelse og inntekt i husstanden, imens kjønnsforskjellene er ukorrelert med BNP.

Da jeg så på om effekten av BNP på interessevariablene innså jeg at det kan ha oppstått et multikollinearitetsproblem mellom BNP og *income*. Dette er fordi de ser ut til å være innbyrdes avhengige av hverandre og følger hverandres trender. Problemet er ikke så overraskende når man tenker seg om, det gir mening at land med høyere BNP har et høyere nivå på inntekten i husstanden.

```
test (girl_bnp)

( 1)  girl_bnp = 0

      F( 1, 9531) =    0.06
      Prob > F =    0.8081

test (income_bnp)

( 1)  income_bnp = 0

      F( 1, 9531) =  195.49
      Prob > F =    0.0000

test (clsize_bnp)

( 1)  clsize_bnp = 0

      F( 1, 9531) =  590.62
      Prob > F =    0.0000
```

Tabell 11: T-test av interaksjonsledd

6 Oppsummering og konklusjoner

Jeg har i oppgaven sett på hvordan elevprestasjoner påvirkes av kjønn, klassestørrelse og inntekt i husstanden. I tillegg har jeg tatt for oss spørsmålet om det er en korrelasjon mellom effekten av disse tre interessevariablene på leseferdigheter og BNP per capita. Gjennom kapitlet «Empiriske resultater» så jeg det som konklusjonen i oppgaven bygger på: Det er en statistisk signifikant effekt av de tre interessevariablene på leseferdighetene hos 4. klassinger. Jeg gikk så videre for å se på om denne effekten varierte med hensyn på BNP. Da jeg så på effekten av BNP på interessevariablene innså jeg at det oppstod et multikollinearitetsproblem mellom BNP og *income*. Det betyr at jeg ikke kan være sikker på effekten mellom de to variablene.

6.1 Begrensninger og mulige utvidelser

En begrensning ved oppgaven er at den tar for seg relativt mange aspekter ved datamaterialet, altså at den ser på flere interessevariabler istedenfor kun én. Jeg har derfor ikke kunnet ta med flere elementer i analysen som av hensyn til at oppgaven da ville blitt for omfattende. Jeg mener derimot at den analysen som er blitt gjort er grundig nok til å gi en tilstrekkelig konklusjon på den problemstillingen som er presentert.

Et annet aspekt er at dette datamaterialet har et begrenset antall variable som påvirker *read*. Det kan bety at det er en del elev- eller familiekarakteristika som ikke er med i datamaterialet, som kunne hatt betydelig innvirkning på testscoren. Muligheten er altså til stede for at det her foreligger et utelatt variabelproblem. Jeg forutsetter i oppgaven at datamaterialet har intern validitet, som vil si at jeg kan trekke gyldige konklusjoner ut fra dataene som er gitt.

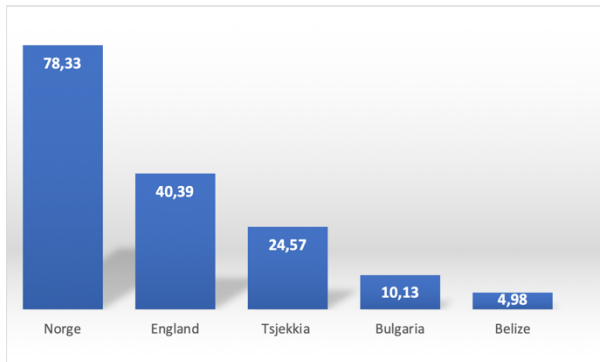
Det er primært tre utvidelser av oppgaven som kunne vært aktuelle:

- Jeg kunne inkludert flere variabler i modellen for å lettere unngå utelatt variabelproblem.
- Jeg kunne gått mer i dybden i datamaterialet til hvert av de fem landene for å se på forskjellige karakteristika innad. Da ville jeg fått et enda bedre bilde på hvorfor et land som feks Bulgaria gjør det bedre enn Norge på lesetesten.
- Jeg kunne også gjort regresjonen med andre regresjonsmetoder for å sammenlikne resultater, eller for å unngå utelatt variabelproblem.

Referanser:

1. Hans Bonesrønning. (2004). Utforming av utdanningspolitikken – Hva kan økonomene bidra med? *Økonomisk forum nr. 3*. Hentet fra <https://samfunnsokonomene.no/wp-content/uploads/2019/05/Trykkutgave-4-2012.pdf>
2. Hanushek, E. A. (2020): Education production functions. I Bradley, S. og Green, C. (red): *Economics of Education*, 2nd Edition, London: Academic Press, 161-170. <http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%202020%20Education%20Production%20Functions.pdf>
3. Stoltenberg, C., Abdelrahman, H. M., Chaudhry, R. A., Fylling, I., Hausstatter, R., Kirkebirkeland, M. A., Lervåg, A., Løken, K. V., Monsen, M., Nereid, C. T., Ogden, T. (2019): *Nye sjanser – bedre læring, kjønnsforskjeller i skoleprestasjoner og utdanningsløp, 14-15*. Hentet fra: <https://nettsteder.regjeringen.no/stoltenbergutvalget/files/2019/02/nou201920190003000dddpdfs.pdf>
4. Sensurveiledning, Hentet fra <https://www.ntnu.no/documents/10455/1263539127/SOS3003+sensurveiledning+V2015.pdf/f82aa1ae-736b-4870-8947-df533c5b49d2>
5. Senter for leseforskning. 2003. *PIRLS – En norsk kortversjon av den internasjonale rapporten om 10-åringers lesekunnskaper*. Norge: Senter for leseforskning. Hentet fra https://learn-eu-central-1-prod-fleet01-xythos.s3-eu-central-1.amazonaws.com/5def77a38a2f7/3638714?response-content-disposition=inline%3B%20filename%2A%3DUTF-8%27%27Solheim_T%25C3%25B8nnesen_En%2520norsk%2520kortversjon%2520av%2520den%2520internasjonale%2520rapporten%2520om%252010-%25C3%25A5ringers%2520lesekunnskaper_2003.pdf&response-content-type=application%2Fpdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20200502T103515Z&X-Amz-SignedHeaders=host&X-Amz-Expires=21600&X-Amz-Credential=AKIAZH6WM4PLYI3L4QWN%2F20200502%2Fau-central-1%2Fs3%2Faws4_request&X-Amz-Signature=f0c6d56eb36f74515aef492edf067c2cd59b86329934829ad8e4dcf65dbc1a7a
6. Thomas, R. L. (2005): *Using statistics in economics*. McGraw-Hill

Appendiks



Figur 8: BNP per capita i tusen USD

Source	SS	df	MS	Number of obs	=	9,541
Model	8325472.01	5	1665094.4	F(5, 9535)	=	159.12
Residual	99779237.1	9,535	10464.5241	Prob > F	=	0.0000
Total	108104709	9,540	11331.7305	R-squared	=	0.0770
				Adj R-squared	=	0.0765
				Root MSE	=	102.3

read	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
girl	18.35209	2.094867	8.76	0.000	14.2457 22.45848
clsize	-.5986367	.173536	-3.45	0.001	-.9388042 -.2584692
income	5.766691	.6137614	9.40	0.000	4.563588 6.969794
par_emp	-26.07427	1.328913	-19.62	0.000	-28.67923 -23.46932
teacher_exp	1.010871	.0993338	10.18	0.000	.8161554 1.205586
_cons	536.0029	5.749655	93.22	0.000	524.7324 547.2735

Tabell 12: Regresjon av modell (3)

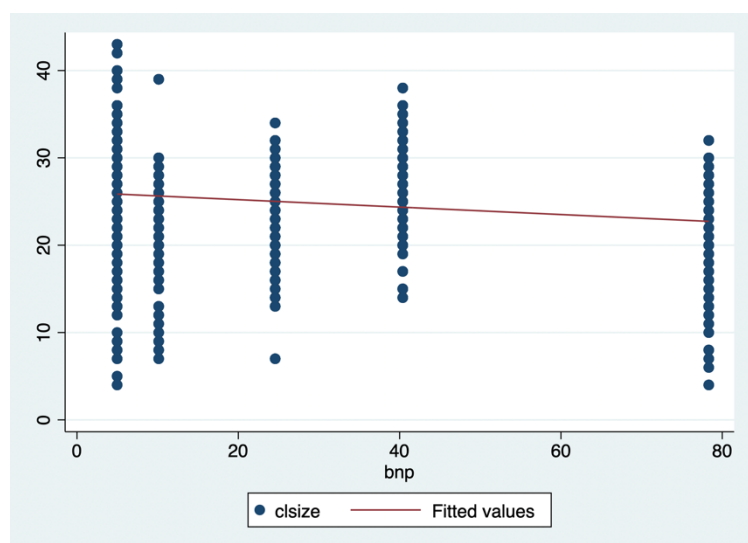
```

test girl
( 1)  girl = 0
      F( 1, 9535) = 76.75
      Prob > F = 0.0000

test income
( 1)  income = 0
      F( 1, 9535) = 88.28
      Prob > F = 0.0000

test clsize
( 1)  clsize = 0
      F( 1, 9535) = 11.90
      Prob > F = 0.0006
    
```

Tabell 13: T-test av interessevariabler



Figur 9: Kryssplott av bnp mot *clsize* og en enkel regresjonslinje

```
test income_sq

( 1) income_sq = 0

      F( 1, 10519) = 18.66
      Prob > F = 0.0000

test clsize_sq

( 1) clsize_sq = 0

      F( 1, 10519) = 823.31
      Prob > F = 0.0000
```

Tabell 14: T-test av kvadratiske ledd

