Bacheloroppgave i samfunnsøkonomi

Sammenheng mellom leseferdigheter og familieinntekt i 4 forskjellige land med forskjellig grad av innvandring og størrelse på familieinntekt blant elever på fjerde trinn.

Joachim Langeland og Petter Ihlebæk

15. mai 2020

Innholdsfortegnelse

Oppsummering

I denne oppgaven har vi undersøkt og analysert hva som påvirker leseferdighetene til elever i fjerdeklasse i fire forskjellige land. Funnene baserer seg på datasettet fra undersøkelsen PIRLS 2001 hvor 35 land deltok. Landene vi har gått ut ifra er:

- Norge (lav innvandring, høy inntekt)
- Belize (lav innvandring, lav inntekt)
- Canada (høyere innvandring, høy inntekt)
- Bulgaria (høy innvandring, lav inntekt)

Grunnen til at det akkurat er disse fire landene som er valgt er fordi alle har forskjellige innvandringsandel og gjennomsnittlig familieinntekt. Det vil tas hovedfokus på om familieinntekt har stor påvirkningskraft til leseferdigheter. Det har også vært fokus på blant annet

- studentens/foreldres innvandringsbakgrunn
- foreldres utdanning
- nivå på leseferdigheter i tidlig alder
- kjønn

Måten det er analysert på er ved hjelp av estimering av variablene gitt i datasettet til PIRLS 2001. Regresjonsanalyse har derfor vært det viktigste hjelpemiddelet. Vi har gjennomført tester for blant annet korrelasjon mellom kontrollvariabler.

Hvordan leseferdighetene påvirkes er et interessant tema, da vi kan finne ut av hva som er med på å fremme bedre prestasjoner og hva som er med på dempe disse prestasjonene. Ved å komme med konklusjoner i denne oppgaven har vi fått et bedre bilde på om det er hold i om innvandringsbakgrunn er med på å dempe prestasjoner, eller om familieinntekt er med på å øke prestasjonene. Vi har tatt variablenes signifikans i betraktning.

Vi hevder at det er god grunn til å påstå en positiv sammenheng mellom familieinntekt og leseferdigheter. Analysen vår vil vise flere sider av saken, men til slutt argumentere for positiv sammenheng. Dette sier på generelt grunnlag på tvers av de fire landene vi har valgt ut.

Landene i rapporten

I datasettet til PIRLS er Belize oppgitt som ett land med lav innvandring, men også ett land med lav familieinntekt
Norge er oppgitt som ett land med lav innvandring og høy familieinntekt
Bulgaria er oppgitt som ett land med høy innvandring og lav familieinntekt.
Det var vanskelig å finne ett land med like høy innvandring som Bulgaria, men som også hadde høy gjennomsnittlig familieinntekt. Det vil derfor være et land som har middels høy innvandring sammenlignet med Bulgaria. Landet som vil være oppgitt er Canada

Med landene oppgitt og størrelsen på innvandring og familieinntekt til de forskjellige landene, vil det være lettere å se på om de faktorene har en påvirkning eller ikke, og i så fall i hvilken grad.

Andre studier

Norske rapporter

Det finnes fra før mye analyser og stoff fra disse temaene om prestasjoner og leseferdigheter på barneskolenivå. Blant annet finnes det flere norske rapporter om dette. En rapport som er en god kortversjon av datasettet til PIRLS 2001 er fra Solheim og Tønnesen[1]. Denne rapporten baserer seg på de tre nordiske landene som var med i dette datasettet, Island, Norge og Sverige. Denne rapporten tar for seg en del faktorer, som f.eks.

- Viktigheten av tidlig lesing i hjemmet
- Forskjellen på alder
- Forskjellen på kjønn
- Om det snakkes landets språk i hjemmet

Dette er bare et par eksempler av faktorene som er gitt i Solheim og Tønnesens kortversjon rapport. Videre har Solheim og Tønnesen en god og utfyllende rapport om resultatene i Norge som tar for seg de fleste faktorer[2]. Denne viser en helhetlig rapport, men har ikke noe sammenligning med andre land, noe som vil gjøres her. dette er bare ett par eksempler av norske rapporter som finnes på nett. Dette gir oss en god inspirasjon på hvordan dette kan gjøres å settes opp.

Utenlandske rapporter

det finnes også mange utenlandske rapporter om datasettet fra PIRLS 2001[3]. en av disse som er den offisielle rapporten om PIRLS 2001 skrevet av Boston college i USA. dette er den offisielle så denne er den mest utfyllende som er å finne. Denne inneholder alle land som var med, med alle variabler, og er på totalt 375 sider inkludert alt av kildehenvisninger, apendix osv. Men det finnes ikke noen direkte sammenligninger mellom land, måten de gjør dert på er å sammenligne øvre del og nedre del av land på hver faktor som testes.

---

[1] kortversjon av PIRLS2001: 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=20200426T115400Z&X-Amz-SignedHeaders=host&X-Amz-Expires=21600&X-Amz-Credential=AKIAZH6WM4PLYI3L4QWN%2F20200426%2Feu-central-1%2Fs3%2Faws4_request&X-Amz-Signature=edeff01b4aeeecc74e1fe4fee0d4f2d60b529a8be522c4f87942715b3ac169fa

[2] utfyllende rapport om PIRLS 2001 Norge:https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/5/pirls_norsk_del_rapport.pdf

3 Internasjonal PIRLS2001 rapport.https://timssandpirls.bc.edu/pirls2001i/pdf/p1_IR_book.pdf

Teoretisk

Metoder som vil bli brukt

Regresjonsanalyser

I oppgaven vil det i hovedsak brukes enkle, multiple og log-lin regresjonsanalyser og hypotesetester. Enkle vil brukes færre ganger enn multiple, men vil brukes for å vise viktigheten av flere variabler og log-lin vil brukes får å se på den prosentvise endringen
Enkle regresjonsanalyser er av formen:
$\hat{y}= \alpha+\beta x_i+\varepsilon_i$
Enkle regresjonsanalyser vil ofte ha et stort restledd og det vil derfor være mer interessant å se på multiple regresjonsanalyser med flere variabler.
Multiple regresjonsanalyser har formen:
$\hat{y}=\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+.....\beta_n x_{ni}$. Multiple vil ha et mindre restledd, og være mer nøyaktige, kontra den enkle regresjonsanalysen.
For å teste om det er grunnlag til å påstå at hver enkelt variabel har påvirkningskraft vil det brukes hypotesetester.

Deskriptiv statistikk

Variabler som vil brukes:

Variabel	**forklaring**	
Idcntry	Land	Landets ID tall
Read	Lesescore/leseferdigheter	Lesescore angitt i poeng
Not_born	elev født i utlandet eller født i Norge	1: elev født i landet 0: elev ikke født i landet
Par_not_born	Foreldre født i utlandet eller født i Norge	1: Foreldre ikke født i landet 0: Foreldre født i landet
Income	Familie inntekt	1-6 1: mindre enn $20,000 2: $20,000-$29,999 3: $30,000-$39,999 4: $40,000-$49,999 5: $50,000-$59,999 6: mer enn $60,000
Par_edu	Foreldres utdanning	1-5 1: universitets grad 2: 3: videregående 4: grunnskole 5: ikke bestått grunnskole
Early_ability	Hvor gode leseferdigheter er i tidlig alder	1-4 1: null ferdighet 2: ferdigheter i liten grad 3: moderate ferdigheter 4: veldig gode ferdigheter
girl	Jente eller gutt	1: jenter 0: gutt

Nedenfor ser vi deskriptiv statistikk til de gitt landene. Hvor gjennomsnitt, standardavvik, minimium observasjon og maksimum observasjon inngår.

Bulgaria

Ut ifra tabellen er Bulgaria et land med meget lav inntekt. Lavest blant disse fire landende med gjennomsnott på kun 1.2 i inntekt. Samtidig ser vi også at innvandring eller barn som ikke er født i landet også er lav, omtrent 7,5%. Det vil derfor være et land med godt utgangspunkt for både påvirkningskraften til lav inntekt og lav innvandring kombinert. Lesescoren ligger på rundt 555 i gjennomsnitt, med et standardavvik i nærheten av 76, minimumscore på 249 og maksimumscore på 743.Bulgaria har høyest gjennomsnitt av disse fire landene.

Variable	Obs	Mean	Std. Dev.	Min	Max
read	3,460	554.9599	76.09892	249.6368	743.7233
idcntry	3,460	100	0	100	100
girl	3,448	.5156613	.4998272	0	1
par_edu	3,249	2.592798	1.085942	1	5
par_not_born	3,380	.0050296	.0707515	0	1
income	3,079	1.276713	.8708402	1	6
not_born	3,339	.0757712	.2646713	0	1
early_abil~y	3,264	2.787071	1.026617	1	4

deskriptiv statistikk Bulgaria

Norge

Ut ifra tabellen er Norge et land med lav innvandring og høy inntekt. Lavere inntekt enn Canada og høyere innvandring enn Bulgaria. Med gjennomsnittlig inntekt på omtrent 4 og innvandringsandel på 9 prosent vil vi se på leseferdigheter i et land med betydelig høyere inntekt enn Bulgaria, men lavere innvandring enn Canada. Lesescoren ligger i nedre sjiktet med 498 i gjennomsnitt. Dette er lavere enn både Canada og Bulgaria, men betydelig høyere enn Belize sine tall. Standardavviket ligger på 78, noe som er et større tall enn Bulgaria og Canada, men mindre enn Belize. Norge er det landet med lavest maksimumscore blant disse landene og ligger nest lavest på minimumscore. Kun Belize ligger under.

Variable	Obs	Mean	Std. Dev.	Min	Max
read	3,459	498.2563	78.36616	228.0606	695.8717
idcntry	3,459	578	0	578	578
girl	3,401	.481035	.4997137	0	1
par_edu	3,098	1.950936	1.055536	1	5
par_not_born	3,374	.0583877	.2345098	0	1
income	2,994	4.064128	1.550755	1	6
not_born	3,355	.0909091	.2875226	0	1
early_abil~y	3,111	2.619415	.9467133	1	4

Deskriptiv statistikk Norge

Canada

ut ifra tabellen er kan vi se at Canada er det landet med høyest inntekt i gjennomsnitt. Det var vanskelig å finne ett land med kombinasjonen høy inntekt og høy innvandring sett i forhold til Belize sine innvandringstall det blir derfor ett land med middels innvandring. Canada er det landet med nest høyest lesescore kun Bulgaria høyere. Men det er det landet med høyest maksimum score på 745 og høyest minimum score. Med ett standardavvik på 69 har det også det laveste standardavviket blant landene.

Variable	Obs	Mean	Std. Dev.	Min	Max
read	8,253	535.9956	69.02971	283.4408	745.4869
idcntry	8,253	124	0	124	124
girl	8,220	.5002433	.5000304	0	1
par_edu	6,681	2.050442	.9491567	1	5
par_not_born	7,300	.1813699	.3853507	0	1
income	6,124	4.628347	1.689817	1	6
not_born	7,748	.2207021	.4147468	0	1
early_abil~y	6,802	2.895766	.9088027	1	4

Deskriptiv statistikk Canada

Belize

Ut ifra tabellen ser vi at Belize er det landet med lavest gjennomsnitt i lesescore. Lesescoren her er på 323 poeng dette er hele 175 til det neste landet. Vi ser også at det er det landet med det største standardavviket med 99 poeng, høyeste poengsum på 710 noe over laveste maksimum score Norge med 695 videre har de også laveste poengsum observert av disse landene. Familie inntekt er det nest lavest kun 0.2 høyere enn Bulgaria. Innvandring er også meget høy med 48% betraktelig høyere enn Canada.

Variable	Obs	Mean	Std. Dev.	Min	Max
read	2,909	323.1901	99.11814	67.90585	710.7631
idcntry	2,909	84	0	84	84
girl	2,765	.5009042	.5000896	0	1
par_edu	1,841	3.625204	1.161285	1	5
par_not_born	2,320	.2788793	.4485447	0	1
income	1,627	1.408728	.9855026	1	6
not_born	2,629	.4880183	.4999515	0	1
early_abil~y	1,854	2.402373	.9839043	1	4

Deskriptiv statistikk Belize

Regresjonsanalyser og hypotesetester

Vi fører 3 ulike regresjonsanalyser hvor vi ved første regresjon utelater alle variabler utenom income, deretter inkluderer vi flere etter hvert. På denne måten vil vi se hvordan lesescoren endrer seg ved at det blir flere variabler i regresjonen. Ved første regresjon vil det være 1 forklaringsvariabel og dette er da hovedforklaringsvariabelen, familieinntekt (income). Dette er for å se hvordan lesescoren endrer seg når vi kun tar hensyn til denne variabelen. Koeffisienten til familieinntekt vil høyst sannsynlig endre seg i takt med at vi inkluderer flere variabler. Dette vil gi innsikt i hvilke faktorer som spiller inn. I andre regresjon vil vi i tillegg ta med variablene girl, og early_ability. I tredje regresjon vil vi ta med alle variablene: girl, early_ability, par_edu, par_not_born, not_born og hovedforklaringsvariabelen income.
Det vil også ses på t-statistikk til variablene, forklaringskraften R-squared og standardfeil. I tillegg til å inkludere seks variablene, vil vi legge til income kvadrert. Dette er fordi vi tror effekten på leseferdigheter vil ta av etter hvert som familieinntekten stiger. Altså en voksende, men avtagende kurve. Vi vil kommentere korrelasjon/interaksjon og gjøre hypotesetester vi mener er relevante for å finne familieinntektens påvirkning.

Regresjonsmodell vil være slik:

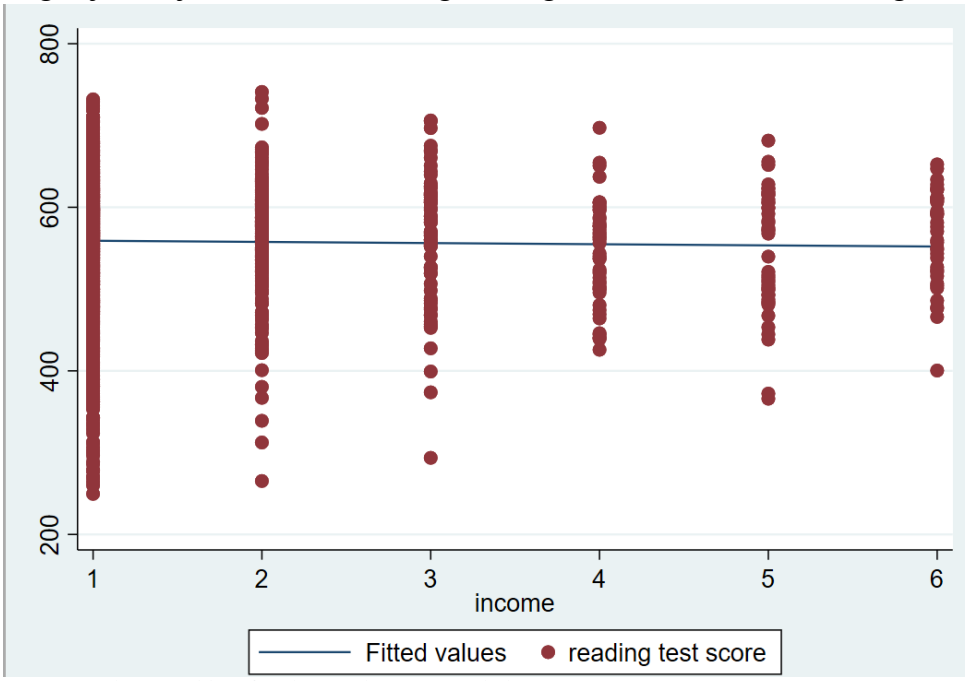
$$read_i = \alpha + \beta_i x_i + \varepsilon_i$$

α - likningens konstantledd
i - tellevariabel
β - koeffisient/stigning. Beskriver hvordan x_i varierer med read.
x - variabelen (income, early_ability etc.).
ε - restledd

Vi tar for oss Bulgaria først. Deretter følger Norge, Canada og Belize til slutt.

Bulgaria

Regresjonslinjen viser at leseferdigheter og familieinntekt er svakt negativt korrelert:



Regresjonslinje med kun familieinntekt som variabel

Vi vil nå gjennomføre regresjonsanalyser med henholdsvis 1, 3 og 6 forklaringsvariabler.

Med 1 forklaringsvariabel

Source	SS	df	MS	Number of obs	=	3,079
				F(1, 3077)	=	0.85
Model	4637.7583	1	4637.7583	Prob > F	=	0.3566
Residual	16789265.8	3,077	5456.37497	R-squared	=	0.0003
				Adj R-squared	=	-0.0000
Total	16793903.5	3,078	5456.10901	Root MSE	=	73.867

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	-1.409553	1.528901	-0.92	0.357	-4.407322	1.588217
_cons	560.482	2.36269	237.22	0.000	555.8494	565.1146

Regresjon 1 forklaringsvariabel

Ved kun familieinntekt som forklaringsvariabel ser vi at lesescoren endrer seg noe. Endringen er negativ med – 1,41poeng per enhets økning i inntekt, og med en standardfeil på 1,5289. Forklaringskraft er på kun 0,03%. Det vil si det er lite i modellen som forteller oss hvorfor leseferdighetene er som de er. Med andre ord er ikke income-koeffisienten alene nok til å kunne gi oss et velbegrunnet resultat.

Med 3 forklaringsvariabler

Source	SS	df	MS	Number of obs	=	3,034
				F(3, 3030)	=	177.92
Model	2459463.54	3	819821.179	Prob > F	=	0.0000
Residual	13961642.5	3,030	4607.80282	R-squared	=	0.1498
				Adj R-squared	=	0.1489
Total	16421106.1	3,033	5414.14641	Root MSE	=	67.881

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	-2.319237	1.432031	-1.62	0.105	-5.127087	.4886129
girl	17.03923	2.486865	6.85	0.000	12.16312	21.91535
early_ability	25.67964	1.220608	21.04	0.000	23.28633	28.07294
_cons	481.6307	4.09148	117.72	0.000	473.6084	489.6531

regresjon med 3 forklaringsvariabler

Vi ser at ved 2 flere forklaringsvariabler minker koeffisienten til familieinntekten. I tillegg ser vi at koeffisientene til girl og early_ability er store positive tall, der girl har en positiv koeffisient på 17. Det vil si at lesescore endrer seg med 17 poeng om du er jente, med en standardfeil på 2,4. Samtidig ser vi at nivået på leseferdigheter i tidlig alder har stor påvirkningskraft med en positiv koeffisient på 25. Det vil si at elever som viser gode leseferdigheter tidlig også gjorde det bedre på 4.trinn. Variabelen har 1,2 i standardfeil. Ved å inkludere disse variablene har modellens forklaringskraft økt betydelig, til 14,98%. Dette er fortsatt ganske lavt.

Vi kan sjekke hvorvidt disse korrelerer, og om korrelasjonen er signifikant. Vi danner interaskjonsvariabelen early_girl = early_ability*girl.

Source	SS	df	MS	Number of obs	=	3,034
				F(4, 3029)	=	133.42
Model	2459810.24	4	614952.559	Prob > F	=	0.0000
Residual	13961295.8	3,029	4609.20959	R-squared	=	0.1498
				Adj R-squared	=	0.1487
Total	16421106.1	3,033	5414.14641	Root MSE	=	67.891

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	-2.317195	1.432269	-1.62	0.106	-5.125512	.4911222
early_ability	26.02422	1.75182	14.86	0.000	22.58934	29.45909
girl	18.90968	7.259336	2.60	0.009	4.675951	33.1434
early_girl	-.6698571	2.442407	-0.27	0.784	-5.458799	4.119085
_cons	480.7079	5.297807	90.74	0.000	470.3202	491.0956

regresjon med interaksjonsledd,

T-statistikken viser –0,27, som er mindre enn den kritiske verdien 1,96. Med andre ord er ikke korrelasjonen statistisk signifikant. Dette gir mening, da det ikke burde ha noe å si hvilket kjønn man er i forhold til hvor tidlig man tilegner seg leseferdigheter. Koeffisientene til early_ability og girl er svekket. Selv om early_girl ikke er statistisk signifikant svekker den forklaringskraften til de to variablene. Dette viser hvordan variabler som ikke er signifikante kan endre de koeffisienten vi ser på.

Med 6 forklaringsvariabler

Source	SS	df	MS	Number of obs	=	2,873
				F(6, 2866)	=	141.92
Model	3291504.78	6	548584.13	Prob > F	=	0.0000
Residual	11078693.2	2,866	3865.55939	R-squared	=	0.2291
				Adj R-squared	=	0.2274
Total	14370198	2,872	5003.55083	Root MSE	=	62.174

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	-1.53491	1.355944	-1.13	0.258	-4.193635	1.123814
not_born	-22.43944	4.502711	-4.98	0.000	-31.26832	-13.61056
par_not_born	-31.19394	18.04044	-1.73	0.084	-66.5675	4.179617
par_edu	-19.47078	1.148642	-16.95	0.000	-21.72303	-17.21854
early_ability	16.75901	1.229846	13.63	0.000	14.34754	19.17048
girl	17.31296	2.342111	7.39	0.000	12.72057	21.90535
_cons	559.481	5.603288	99.85	0.000	548.4941	570.4678

Regresjon med 6 forklaringsvariabler

Ved alle 6 forklaringsvariablene inkludert har vi ett mer riktig bilde av hvordan variablene endrer lesescoren. Her endrer koeffisienten til income seg fra -2,32 opp til -1,53. Standardfeil endres minimalt fra 1,4 til 1,35 ned -0,005.

variablene som var med i forrige regresjon har også endret seg, hvor:

- Girl endret seg fra 17 opp til 17,31 en liten endring sett i forhold til tallet, og standardfeilen endrer seg ned med 0,1 til 2,3.
- Early ability endret seg fra 25,67 ned til 16,75 og standardfeil holder seg likt på 1,2

De nye variablene, not_born, par_not_born og par_edu har alle negative koeffisienter. Hvor:

- not_born har en koeffisient på -22,44, som vil si at lesescoren endrer seg med – 22,44 visst du er født utenlands med en standardfeil på 4,5.
- par_not_born har en negativ koeffsient på -31,19 som vil si at lesescoren endrer seg med -31,19 om foreldrene er født utenlands, her med en standardfeil på 18,4.
- Par_edu har en negativ koeffisient på -19,47. her vil lesescoren endre seg -19,47 per en enhets økning i foreldres utdanning. Her med en standardfeil på 1,1.

Videre ser vi også at modellens forklaringskraft R^2 har endret seg med antall variabler. Forklaringskraften til modellen endres seg og er på 0,03%, 14% og 22% med henholdsvis 1, 3 og 6 forklaringsvariabler. Dette endrer seg i takt med hvor stor t-statistikken er på hver enkelt modell, hvor den første kun har income som forklaringsvariabel med liten t-statistikk (lite signifikant). Dette gir en veldig lav forklaringskraft i modellen.

Vi legger nå til income kvadrert, income_sqr.

Source	SS	df	MS	Number of obs	=	2,873
				F(7, 2865)	=	121.60
Model	3291517.49	7	470216.784	Prob > F	=	0.0000
Residual	11078680.5	2,865	3866.90419	R-squared	=	0.2291
				Adj R-squared	=	0.2272
Total	14370198	2,872	5003.55083	Root MSE	=	62.184

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	-1.87725	6.123807	-0.31	0.759	-13.88476	10.13026
income_sqr	.0589776	1.0288	0.06	0.954	-1.958286	2.076241
not_born	-22.43909	4.503498	-4.98	0.000	-31.26952	-13.60867
par_not_born	-31.15637	18.05548	-1.73	0.085	-66.55942	4.246673
par_edu	-19.4728	1.149377	-16.94	0.000	-21.72648	-17.21911
early_ability	16.75937	1.230076	13.62	0.000	14.34745	19.1713
girl	17.31012	2.343043	7.39	0.000	12.7159	21.90434
_cons	559.7827	7.688132	72.81	0.000	544.7078	574.8575

Regresjon med income kvadrert

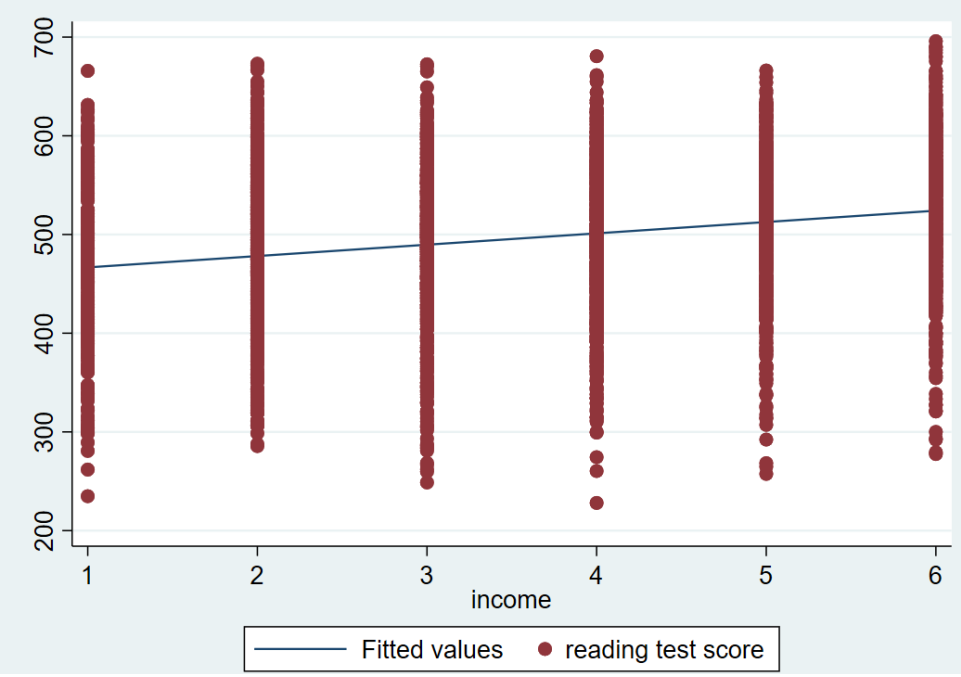
Deriverer vi uttrykket for read med hensyn på income, får vi:
Read = -1,87725+ 2*0,589776*income
La oss sette inn snittet for income, omtrent 1,2:
Read = -1,87725 + 2*0,589776*1,2 = -0,4617
Så setter vi inn income = 4:
Read = -1,87725 + 2*0,589776*4 = 2,8409

I regresjonen fant vi en negativ sammenheng mellom leseferdigheter og inntekt. Det deriverte uttrykket, og utregningene, viser at leseferdighetene blir bedre når inntekten vokser dersom den er høy. Ved gjennomsnittlig inntekt henger økt inntekt og leseferdigheter negativt sammen. Vi vil se om dette også gjelder i Norge, der korrelasjonen er annerledes og inntekten høyere.

Figur 1regresjon med income kvadrert

Norge

I Norge har vi en positiv korrelasjon mellom leseferdigheter og familieinntekt:



Regresjonslinje med income som variabel

Med 1 forklaringsvariabel

Source	SS	df	MS	Number of obs	=	2,994
				F(1, 2992)	=	165.23
Model	950952.506	1	950952.506	Prob > F	=	0.0000
Residual	17219917.2	2,992	5755.31992	R-squared	=	0.0523
				Adj R-squared	=	0.0520
Total	18170869.7	2,993	6071.12252	Root MSE	=	75.864

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	11.49431	.8942074	12.85	0.000	9.740987	13.24763
_cons	455.1808	3.889666	117.02	0.000	447.5541	462.8075

Regresjon med 1 forklaringsvariabel

I Norge ser vi at høyere familieinntekt sannsynligvis har motsatte effekt enn i Bulgaria, der leseferdighetene ble noe svakere i takt med høyere inntekt. I Norge er det sterkt motsatt. De med høyere inntekt kan vise til betydelig bedre leseferdigheter enn de med lav inntekt. Koeffisienten til familieinntekt med kun 1 variabel er her på 11,49. Og en standardfeil på 0,8

Med 3 forklaringsvariabler

Source	SS	df	MS	Number of obs	=	2,921
				F(3, 2917)	=	243.77
Model	3524605.97	3	1174868.66	Prob > F	=	0.0000
Residual	14058448.9	2,917	4819.48883	R-squared	=	0.2005
				Adj R-squared	=	0.1996
Total	17583054.9	2,920	6021.59414	Root MSE	=	69.423

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	10.90344	.8297816	13.14	0.000	9.27642	12.53045
early_ability	30.21465	1.371732	22.03	0.000	27.52499	32.90431
girl	10.70718	2.601293	4.12	0.000	5.60662	15.80774
_cons	373.9805	5.051345	74.04	0.000	364.0759	383.885

Regresjon med 3 forklaringsvariabler

Når vi også her inkluderer flere variabler i modellen ser vi at koeffisienten til familieinntekt endrer seg. Her endrer den seg fra 11,49 til 10,9. Med standardfeil på 0,8

Med 6 forklaringsvariabler

Source	SS	df	MS	Number of obs	=	2,813
				F(6, 2806)	=	162.03
Model	4295534.29	6	715922.381	Prob > F	=	0.0000
Residual	12397809.7	2,806	4418.32135	R-squared	=	0.2573
				Adj R-squared	=	0.2557
Total	16693344	2,812	5936.46657	Root MSE	=	66.47

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	5.013184	.8977603	5.58	0.000	3.252847	6.773521
not_born	-18.13336	4.872532	-3.72	0.000	-27.68747	-8.579247
par_not_born	-29.95809	6.267074	-4.78	0.000	-42.24663	-17.66955
par_edu	-17.34024	1.305888	-13.28	0.000	-19.90084	-14.77964
early_ability	29.15321	1.339998	21.76	0.000	26.52573	31.78069
girl	11.07486	2.539015	4.36	0.000	6.096332	16.05338
_cons	437.9515	6.537319	66.99	0.000	425.1331	450.77

Regresjon med 6 forklaringsvariabler

Med alle variablene tatt med minker income-koeffisienten til 5. Oppe i høyre hjørnet ser vi at R^2, modellens forklaringskraft, er 25,7%. I estimatene med en og tre variabler var forklaringskraften henholdsvis 5,2% og 20%.
Vi legger merke til t-statistikken til hver variabel. For early_ability og par_edu er denne veldig stor. Med andre ord er dette svært signifikante variabler. Utelater vi disse går forklaringskraften drastisk ned. Dette er spesielt synlig fra den første regresjonen til den andre regresjonen med tre variabler.
Siden alle variablene korrelerer med read, må vi sjekke om variablene seg imellom korrelerer direkte eller indirekte. Med indirekte korrelasjon menes det at de korrelerer gjennom den felles variabelen read. Derfor må det vises hvordan variablene korrelerer med income. Dette gjøres gjennom hypotesetester.
Vi tester par_edu og income gjennom interaksjonsvariabelen edu_income = par_edu*income:

Source	SS	df	MS	Number of obs	=	2,813
				F(7, 2805)	=	139.09
Model	4301249.79	7	614464.255	Prob > F	=	0.0000
Residual	12392094.2	2,805	4417.8589	R-squared	=	0.2577
				Adj R-squared	=	0.2558
Total	16693344	2,812	5936.46657	Root MSE	=	66.467

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	6.764626	1.78241	3.80	0.000	3.269658	10.25959
not_born	-18.08129	4.872492	-3.71	0.000	-27.63532	-8.527253
par_not_born	-30.30273	6.274067	-4.83	0.000	-42.60499	-18.00048
par_edu	-13.82734	3.353194	-4.12	0.000	-20.40231	-7.252359
arly_ability	29.13272	1.340049	21.74	0.000	26.50514	31.7603
girl	11.01579	2.539413	4.34	0.000	6.036488	15.9951
edu_income	-.9198282	.8086967	-1.14	0.255	-2.505529	.6658724
_cons	430.7549	9.097503	47.35	0.000	412.9165	448.5934

Regresjon med interaksjonleddet edu_income

Vi ser at med et konfidensintervall på 95% er edu_income ikke signifikant, selv om t-statistikken er på -1,14. Den kritiske verdien er 1,96. Allikevel endrer den koeffisientene foran de andre variablene. Endringene på income, par_edu og early_ability er ganske store. Da ser vi at selv om korrelasjonen edu_income regnes som signifikant påvirker den resultatene våre. Dette er en svakhet ved modellen.

Vi vil nå gjøre en analyse ved å kvadrere income, income_sqr = income^2.

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	-2.851937	4.665507	-0.61	0.541	-11.99986	6.29599
income_sqr	1.883164	.601099	3.13	0.002	.7045551	3.061774
_cons	477.8541	8.213563	58.18	0.000	461.7493	493.9589

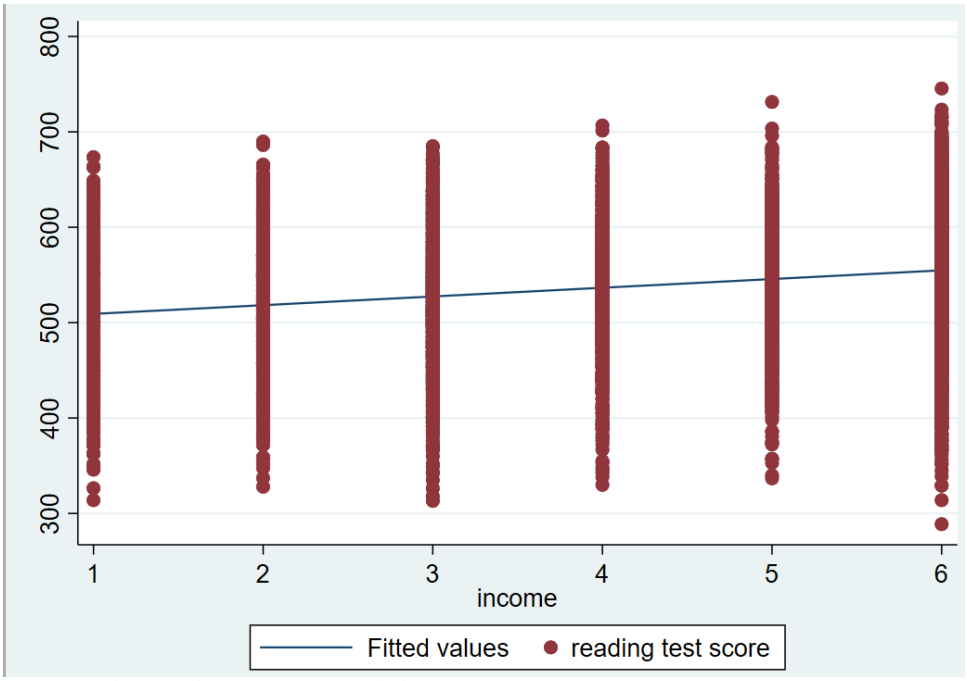
Regresjon med income kvadrert

Deriverer vi uttrykket får vi:
Read = -2,851937 + 2*1.883164*income
Setter inn 3 (under gjennomsnitt) og 5 (over gjennomsnitt).
Read(3) = 9,1509
Read(5) = 15,9797
Veksten er sterkere ved høyere inntekt enn ved lavere inntekt. Vi ser av ligningen at inntekten må være svært lav for at de ikke henger positivt sammen med leseferdigheter.

I Norge ser det altså ut som at vi har en positiv sammenheng mellom familieinntekt og leseferdigheter. Allikevel er det også andre signifikante variabler som påvirker mye. Inntekten alene forklarer lite av forskjellene i leseferdigheter.

Canada

Også i Canada er det en svak positiv korrelasjon:



Regresjonslinje med income som variabel

Med 1 forklaringsvariabel

Source	SS	df	MS	Number of obs	=	6,124
				F(1, 6122)	=	332.49
Model	1457987.03	1	1457987.03	Prob > F	=	0.0000
Residual	26845638.8	6,122	4385.10925	R-squared	=	0.0515
				Adj R-squared	=	0.0514
Total	28303625.8	6,123	4622.50953	Root MSE	=	66.22

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	9.13177	.5008046	18.23	0.000	8.150017	10.11352
_cons	500.0476	2.46753	202.65	0.000	495.2104	504.8848

Regresjon med 1 forklaringsvariabel

Canada har som Norge en positiv sammenheng mellom leseferdigheter og familieinntekt. Her er koeffisienten på 9,1 og standardfeil på 0,5. Forklaringskraft på 5%, likt som i Norge.

Med 3 forklaringsvariabler

Source	SS	df	MS	Number of obs	=	6,035
				F(3, 6031)	=	375.78
Model	4399890.52	3	1466630.17	Prob > F	=	0.0000
Residual	23538250.9	6,031	3902.87696	R-squared	=	0.1575
				Adj R-squared	=	0.1571
Total	27938141.5	6,034	4630.11957	Root MSE	=	62.473

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	8.894295	.4788785	18.57	0.000	7.955522	9.833068
early_ability	23.1251	.8907077	25.96	0.000	21.379	24.87121
girl	10.77563	1.618289	6.66	0.000	7.603208	13.94806
_cons	428.8839	3.498733	122.58	0.000	422.0251	435.7427

Regresjon med 3 forklaringsvariabler

Som vi så på de andre regresjonene vil koeffisienten bli mindre på andre regresjon, når early_ability og girl blir inkludert. Koeffisienten går ned kun 0,2 poeng, som er en liten endring. De nye variablene early_ability og girl har koeffisienter på henholdsvis 23 og 10. Med disse nye variablene vil vi ha en ny forklaringskraft på 15,75%. En stor endring på omtrent 10%. Samme utvikling som i Norge.

Med 6 forklaringsvariabler

Source	SS	df	MS	Number of obs	=	5,214
				F(6, 5207)	=	262.61
Model	5513198.66	6	918866.443	Prob > F	=	0.0000
Residual	18219462	5,207	3499.03245	R-squared	=	0.2323
				Adj R-squared	=	0.2314
Total	23732660.6	5,213	4552.59172	Root MSE	=	59.153

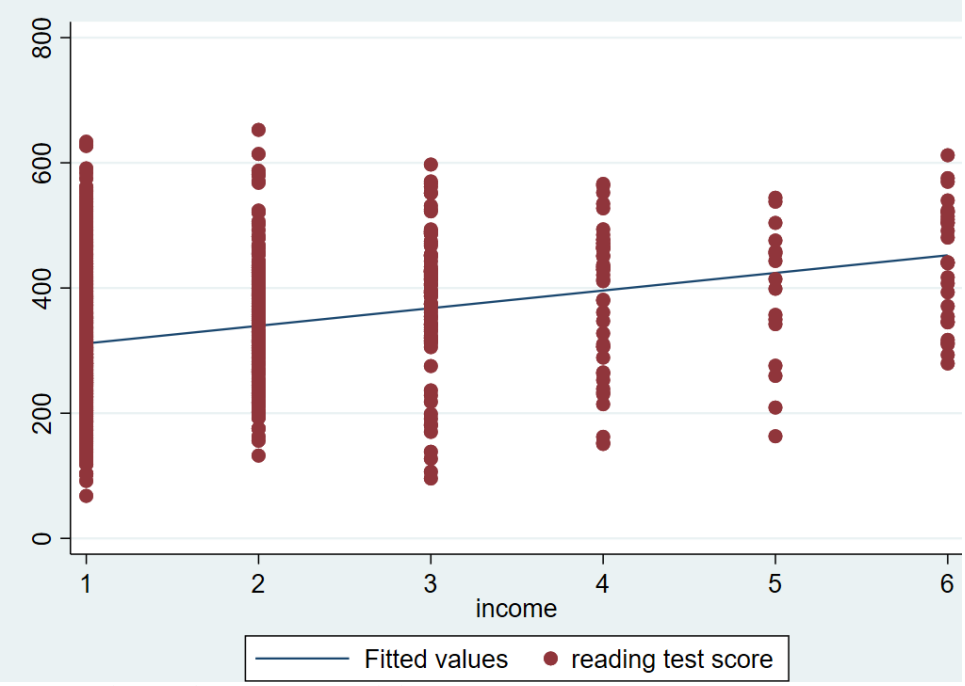
read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	4.535525	.5538026	8.19	0.000	3.44984	5.621211
not_born	-37.74514	2.146836	-17.58	0.000	-41.95384	-33.53644
par_not_born	.4616685	2.34886	0.20	0.844	-4.143083	5.06642
par_edu	-13.57641	.9595801	-14.15	0.000	-15.45759	-11.69523
early_ability	21.98515	.9184615	23.94	0.000	20.18458	23.78572
girl	10.4673	1.651726	6.34	0.000	7.229222	13.70537
_cons	489.9761	4.904589	99.90	0.000	480.3611	499.5912

Regresjon med 6 forklaringsvariabler

Med alle variablene inkludert ser vi også her som ved Norge en halvering av koeffisienten til income. De nye koeffisientene not_born og par_edu har begge negative koeffisienter på henholdsvis –37.7 og –13.6, med t-statisitkk på -17 og –14. Den nye variabelen par_not_born har en meget lav koeffisient på 0,46. Ny forklaringskraft ved 6 variabler viser samme utvikling som Norge: fra 5% til 15% til 23%.

Belize

Belize har en positiv korrelasjon mellom leseferdigheter og familieinntekt.



Regresjonslinje med income som variabel

Med 1 forklaringsvariabel

Source	SS	df	MS		
Model	1252976.81	1	1252976.81	Number of obs	= 1,627
Residual	14497842.3	1,625	8921.74913	F(1, 1625)	= 140.44
Total	15750819.1	1,626	9686.85064	Prob > F	= 0.0000
				R-squared	= 0.0795
				Adj R-squared	= 0.0790
				Root MSE	= 94.455

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	28.16784	2.376878	11.85	0.000	23.50577	32.82991
_cons	283.2883	4.085972	69.33	0.000	275.274	291.3026

Rregresjon med 1 forklaringsvariabel

Vi ser også her ved Belize at koeffisienten til Income er positiv, men her mye høyere en både Canada og Norge, med en koeffisient på 28 og standardfeil på 2,3 vil det være stor feil i modellen. med en t-statistikk på 11.85. Forklaringskraften R^2 er til sammenligning den største ved første regresjon blant landene med 7,95%

Med 3 forklaringsvariabler

Source	SS	df	MS		
Model	2604599.93	3	868199.975	Number of obs	= 1,525
Residual	12403224.9	1,521	8154.65149	F(3, 1521)	= 106.47
Total	15007824.8	1,524	9847.6541	Prob > F	= 0.0000
				R-squared	= 0.1735
				Adj R-squared	= 0.1719
				Root MSE	= 90.303

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	24.28749	2.347529	10.35	0.000	19.68275	28.89223
early_ability	27.50042	2.395895	11.48	0.000	22.80081	32.20002
girl	19.25241	4.662103	4.13	0.000	10.10758	28.39724
_cons	214.7781	6.7364	31.88	0.000	201.5645	227.9917

Regresjon med 3 forklaringsvariabler

Her ved 2 nye variabler ser vi at koeffisienten har samme utvikling som de andre landene, med en nedgang her på -4 ned til 24. Standardfeil holder seg noenlunde likt på 2.35.
De nye variablene har til sammenligning høye koeffisienter på henholdsvis 29 og 19.
forklaringskraften R^2 har gått opp til 17,35%

Med 6 forklaringsvariabler

Source	SS	df	MS	Number of obs	=	1,208
				F(6, 1201)	=	67.76
Model	3147500.94	6	524583.49	Prob > F	=	0.0000
Residual	9297216.64	1,201	7741.22951	R-squared	=	0.2529
				Adj R-squared	=	0.2492
Total	12444717.6	1,207	10310.4537	Root MSE	=	87.984

read	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	15.07196	2.721237	5.54	0.000	9.733053	20.41087
not_born	-22.30529	5.30041	-4.21	0.000	-32.70438	-11.9062
par_not_born	8.093505	5.841559	1.39	0.166	-3.36729	19.5543
par_edu	-22.04939	2.420635	-9.11	0.000	-26.79853	-17.30024
early_ability	24.25373	2.704256	8.97	0.000	18.94813	29.55932
girl	22.14502	5.12283	4.32	0.000	12.09433	32.19571
_cons	323.3623	13.79015	23.45	0.000	296.3069	350.4178

Regresjon med 6 forklaringsvariabler

Med alle variablene inkludert ser vi her at income-koeffisienten har gått ned ytterligere, til 15. Standardfeil er på 2,7.
De variablene som var inkludert på forrige regresjon har ulik utvikling. Early_ability har en ny koeffisient på 24, ned 4, og girl med ny koeffisient på 22, opp omtrent 3.
De nye variablene har for det meste store negative koeffisienter der par_not_born skiller seg ut med positiv koeffisient. Not_born og par_edu har begge negative koeffisenter på -22.

Vi ser altså sterk positiv sammenheng mellom familieinntekt og leseferdigheter i Belize.

Konklusjon

Etter regresjonsanalysene vi har gjort har vi kommet fram til at det er stor sammenheng mellom leseferdigheter og familieinntekten, der 3 av landene er sterkt positivt korrelert, mens det ene landet er svakt negativt korrelert. Den negative korrelasjonen fant vi i Bulgaria, men denne var avtagende og etter hvert voksende. Dette er en meget begrenset modell på grunn av variabler vi har valgt ut og det kan derfor være nyttig å inkludere flere variabler. Det vil også være interessant å se på korrelasjonen mellom flere av variablene. I tillegg kunne vi ha inkludert flere land, eller en annen sammensetning av land.
Dette ville forbedret modellen og gjort oss sikrere, eller kanskje mindre sikre, i vår sak.