SØK2901 Bacheloroppgave

Hjemmet som forutsetning for leseferdigheter

En økonometrisk analyse av antall bøker i husholdningen og deres effekt på leseferdigheter

---

Christoffer Andersen

Emil Karlsen

Jakob Oseid

NTNU
Fakultet for økonomi
Institutt for samfunnsøkonomi

Vår, 2020

Sammendrag

Denne oppgaven utforsker de bakenforliggende årsakene til at enkelte elever scorer høyere på leseferdigheter. Vi isolerer familiekarakteristika som den klart viktigste faktoren i elevers leseferdigheter, og ser et stort antall bøker i hjemmet som en indikator på høy sosioøkonomisk status. Denne konklusjonen er forankret i en regresjonsanalyse som benytter relevante variabler for familiekarakteristika mot PIRLS-undersøkelsens mål på leseferdigheter. I tillegg undersøkes det om tilgang på PC i klasserommet påvirker leseferdigheten til elevene. Vår analyse finner ikke noe grunnlag for å hverken bekrefte eller avkrefte hypotesen om at pc har noen innvirkning på elevers leseferdigheter. Videre konkluderes det med at et høyere antall bøker i hjemmet gir bedre leseferdigheter, men at denne effekten ikke kan ses isolert fra øvrig familiekarakteristika da det ikke nødvendighvis er antallet bøker i hjemmet i seg selv som er utslagsgivende, men heller de sosioøkonomiske forutsetningene til eleven og bøker i hjemmet forstås som en indikator på dette.

Abstract

The thesis explores the underlying factors of why some students achieve higher scores in reading ability tests. We isolate family-characteristics as the predominant explanatory factor when looking at differences in reading ability between students, and observe that a high density of books in the home is often a symptom of high socioeconomic capital. We utilized data from the PIRLS test undertaken in Norway in 2001, in our analysis. In addition to looking at the family characteristics, the thesis also investigates the potential benefit of having access to PCs in the classroom on students reading ability. However, we cannot find any significant evidence of either or in or in our analysis of the model. The thesis finds evidence for a relationship between books at home and student reading ability. However, one has to be careful when interpreting these results as more books in itself might not be the explanatory factor but rather a symptom of other causality, mainly the families' socioeconomic status. We thus argue that books at home might be a good indicator of socioeconomic status.

Skriveperiode

01. 01. 2020 – 15. 05. 2020

Veileder

Prof. Bjarne Strøm, Institutt for samfunnsøkonomi, NTNU.

Innhold

1 Innledning og problemstilling

Mye tyder på at dynamikken i hjemmet har stor innvirkning på elevers prestasjoner på skolen. I deres artikkel om hva som skaper forskjeller i prestasjon *mellom* skoler, argumenterer blant annet Turmo og Lie (2004) at det i stor grad er elevers sosioøkonomiske bakgrunn som skaper ulikheter i elevprestasjoner. Elevers utgangspunkt for prestasjon er dermed sterk knyttet til hjemmets økonomiske, kulturelle og sosiale kapital. Utdanning isoleres i flere studier som en utslagsgivende sosioøkonomisk faktor, der det argumenteres for at høyt utdannede foreldre, ofte har mange bøker tilgjengelig i hjemmet, samt høyere inntekt. Dette viser seg å ha utslagsgivende effekt på deres barns leseferdigheter. Særlig har barns interaksjon med bøker *før* de begynner på skolen stor effekt på deres prestasjoner på senere lesetester (Gustafsson mfl., 2011; Myrberg & Rosén, 2009). Vi bruker bøker i hjemmet som en indikator på sosioøkonomisk stand. I vårt argument er sosioøkonomiske ulikheter, herunder familiekarakteristika, essensielt for elevers leseferdigheter.

Vår oppgave har som mål, gjennom økonometrisk analyse, å undersøke hvordan leseferdigheter påvirkes av antall bøker i hjemmet, ved bruk av andre relevante variabler som tidlig leseevne, foreldres utdanning, familieinntekt og tilgang på PC i klasserommet, med bakgrunn i følgende problemstilling:

Har antall bøker i hjemmet en utslagsgivende effekt på fjerdeklassingers leseferdigheter, og er tilgang til skjerm i skolehverdagen en distraksjon eller en ressurs i utviklingen av disse ferdighetene?

Det empiriske grunnlaget for analysen vil være Progress In International Reading Literacy Study (PIRLS) undersøkelsen, fra 2001. Basert på datasettet konstruerer vi tre modeller som benytter minste kvadratersmetode (OLS-regresjon), henholdsvis en grunnmodell som betrakter den direkte sammenhengen mellom leseferdighet og bøker i hjemmet, deretter en utvidet modell som inneholder de ovennevnte kontrollvariablene inntekt, språk i hjemmet, foreldres utdannelse og tilgang på PC i klasserommet. I tredje modell inkluderer vi elevens tidligere leseevne samt om de er født i Norge eller ikke.

Etter denne innledningen i del 1, vil vi videre presentere vårt teoretiske grunnlag i del 2. Del 3 presenterer våre variabler, og hvordan vi har bearbeidet dem, før vi i del 4 presenterer vår økonometriske strategi og hypotesetester. I del 5 legger vi frem vi våre resultater, fulgt av del 6 der vi diskuterer disse resultatene. Del 7 vil bestå av en diskusjon av våre funn, før vi konkluderer oppgaven.

2 Teori og teoretiske utgangspunkt

2.1 Regresjon

Dette segmentet inneholder en innledende redegjørelse for regresjon; minstekvadraters metode; determinasjonskoeffesienten R^2 som mål på predikasjonskraft; og en kort redegjørelse for multippel regresjon. Deretter følger en redegjørelse for skoleproduktfunksjonen, fulgt av oppgavens teoretiske utgangspunkt for problemstillingen. Tilpasning av skoleproduktfunksjonen i tråd med våre variabler, gjøres etter del 3, deskriptiv statistikk.

2.1.1 Grunnleggende to-variabel korrelasjon og regresjon

Denne oppgaven skal først undersøke et forhold mellom leseferdighet som avhengig variabel og bøker i hjemmet som forklaringsvariabel. Antagelsen er at dette forholdet er lineært og kan uttrykkes som:

$$E(y) = \alpha + \beta x \tag{2.1}$$

Dette er generelt, men i vår oppgave er $E(y)$ *forventet* leseferdighet for en gitt mengde bøker i hjemmet x. α og β er ukjente populasjonsparametere, og representerer henholdsvis skjæringspunktet og stigningstallet til den rette linjen som den lineære sammenhengen uttrykker. Faktisk leseferdighet, Y, er ikke alltid den samme som forventet leseferdighet $E(y)$ og forskjellen mellom de to kan forstås som avvik og uttrykkes som ϵ og kalles et stokastisk restledd:

$$Y = E(y) + \epsilon \tag{2.2}$$

Setter inn (2.1) i (2.2) som gir:

$$Y = \alpha + \beta x + \epsilon \tag{2.3}$$

Den faktiske regresjonslinjen, eller populasjonsmodellen, er ukjent, men vi bruker data for x og y for å estimere α og β. Hvis predikert verdi uttrykkes som $\hat{Y}$, og a og b er et anslag for henholdsvis α og β, så får vi dermed:

$$\hat{Y} = a + bx \tag{2.4}$$

$$Y = \hat{Y} + e \tag{2.5}$$

Der e er residualen mellom faktisk og predikert verdi på observasjonen, og vil være det empiriske motstykket til det uobserverte stokastiske restleddet ϵ. Utfordringen ved regresjon ligger i å beregne verdier på a og b slik at regresjonslinjen best passer observasjonene, noe som intuitivt gjøres ved å velge a og b slik at residualene blir minimert. Noen av disse residualene vil være positive, mens andre vil være negative. Disse kvadreres, derav metodens navn.

2.1.2 Minste kvadraters metode (Ordinary least squares, OLS)

For å se på effekten av bøker i hjemmet på leseferdighetene til en gjennomsnittlig elev, benytter vi oss av minste kvadraters metode (OLS). Denne regresjonsformen estimerer et forhold mellom en avhengig variabel Y, og en eller flere uavhengige variabler X. En sammenheng predikeres på bakgrunn av minimering av summen mellom de observerte verdiene av den avhengige og de uavhengige variablene. Dette føyer en rett linje til data, hvor linjen minimerer summen av kvadrerte residualer. Vi får altså en lineær sammenheng mellom x og y. En enkel lineær regresjon kan skrives på denne formen:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, 2, 3, ..., n. \tag{2.6}$$

Y er den predikerte verdien til den avhengige variabelen, α er konstantleddet til regresjonslinjen, og β er det estimerte stigningstallet til regresjonslinjen. Estimasjonen vil bare være en tilnærming til den faktiske sammenhengen mellom y og x, som nødvendiggjør en variabel for å fange opp uforklart sammenheng og variasjon ϵ, altså det stokastisk restledd. I en to-variabel klassisk regresjonsmodell gjøres det en rekke antagelser for forklaringsvariabelen x; 1) Den må være ikke-stokastisk (ikke tilfeldig), 2) forventningsverdien av estimeringsfeil, samt varians i forventningsverdien kvadrert, må være lik null, og 3) estimeringsfeilen må være normalfordelt. Deretter er dette et optimeringsproblem med mål om å minimere estimeringsfeil, som forklart i 2.1. Summen av kvadrater kan skrives som:

$$S = \Sigma e_i^2 = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma(Y_i - a - bx_i)^2 \tag{2.7}$$

Der summene er over alle i, altså observerte verdier, i vårt tilfelle elever og S, forstås som summen av kvadrerte residualer. Deretter er dette et rent optimeringsproblem, som settes lik null. Dette kalles ofte normalligningene:

$$\frac{\delta S}{\delta a} = -2\Sigma(Y_i - a - bx_i) = 0 \tag{2.8}$$

$$\frac{\delta S}{\delta b} = -2\Sigma x_i(Y_i - a - bx_i) = 0 \tag{2.9}$$

Om vi tilegner summen n-antall observasjoner innebærer ligning (2.8) at:

$$\begin{aligned} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} bx_i &= 0 \\ \sum y_i - n \cdot a - b \sum x_i &= 0 \end{aligned}$$

$$\rightarrow \quad a \quad = \quad \frac{1}{n} \sum y_i - b \frac{1}{n} \sum x_1$$

Hvor om $\frac{1}{n} \sum y_i$ uttrykkes som $\bar{y}$ og $b\frac{1}{n} \sum x_1$ uttrykkes som $\bar{x}$ gir dette:

$$a = \bar{y} - b\bar{x} \tag{2.10}$$

Vi har da funnet a, som er *estimator* for konstantleddet (α). Videre kan dette løses for b ved å reformulere residualkvadratsummen:

$$S = \Sigma(Y_i - a - bx_i)^2 \tag{2.11}$$

Om vi setter inn ligning (2.10) har vi da at:

$$S' = \Sigma(y_i - (y_i - b\bar{x}) - bx_i)^2 \tag{2.12}$$

$$S' = \Sigma(y_i - \bar{y}) - b(x_i - \bar{x}) \tag{2.13}$$

og minimerer S' med hensyn på b:

$$\frac{\delta S}{\delta b} = 2\Sigma[(y - \bar{y}) - b(x_i - \bar{x})] \cdot -(x_i - \bar{x}) = 0 \tag{2.14}$$

Som gir løsning for stigningstallet b:

$$b = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2} \tag{2.15}$$

Altså er likning (10) $a = \bar{y} - b\bar{x}$ OLS-*estimator* for α
og likning (15) $b = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$ OLS-*estimator* for β

Vi har data for x_i og y_i og kan følgelig beregne a og b som vil gi et uttrykk $\hat{y} = a + bx$,der $\frac{\delta \hat{y}}{\delta x} = b$ tolkes til at økning av x med 1 enhet gir predikert økning i y med b enheter.

2.1.3 Determinasjonskoeffesienten R^2, eller lettere sagt: hvor "god" er den empiriske modellen?

R^2 er et mål på modellens forklaringskraft, og beregnes intuitivt som at:

total sum av kvadrater (SST) = forklart sum av kvadrater (SSE) + residuale sum av kvadrater (SSR)

Dette kan uttrykkes til å forstå R^2 for å forklare variasjon tilskrevet x (SSE), delt på total variasjon i y (SST), ($\frac{SSE}{SST}$), og vi kan følgelig uttrykke at:

$$R^2 = \frac{SSE}{SST} = \frac{SST - SSR}{SST} = 1 - \frac{SSR}{SST} \tag{2.16}$$

Følgelig er R^2 1 minus residualvariasjonens andel av total variasjon. Er $R^2 = 1$ forklarer modellen *all* variasjon i y og vil ha en "sterk" forklaringskraft da alle observasjonene vil ligge på regresjonslinjen. Er $R^2 = 0$ forklarer modellen ingen variasjon i y. R^2 er dermed et utrykk for en models forklaringskraft.

2.1.4 Multippel regresjon

Likning (2.6) kan utvides til flere forklaringsvariabler:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \qquad i = 1, 2, 3, ..., n. \tag{2.17}$$

Gitt en forutsetning om at ingen av variablene x_{ni} er kolineære, er fremgangen intuitivt sett i hovedsak den samme, da målsetningen blir igjen å minimere summen av kvadrater. Det er altså forståelsesmessig likt, men som Thomas (2005, s. 389) uttrykker: *"it's just that the algebra is more cumbersome."* og vi kan vise til estimatorene for b_1. b_2 og b_3 i en hypotetisk modell med tre variabler for dette, for å illustrere:

$$b_1 = \bar{y} - b_2\bar{x}_2 - b_3\bar{x}_3 \tag{2.18}$$

$$b_2 = \frac{\Sigma y x_2 \Sigma_3^2 - \Sigma y x_3 \Sigma x_2 x_3}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2} \tag{2.19}$$

$$b_3 = \frac{\Sigma y x_3 \Sigma x_2^2 - \Sigma y x_2 \Sigma x_3 x_2}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2} \tag{2.20}$$

Dette hadde riktignok sett ryddig ut i vektor-matrise-form, men matriser er forbi nivået som forventes for denne oppgaven.

2.1.5 Hypotesetesting

2.1.5.1 F-test

Vi vil benytte oss av en F-test for vår hypotesetesting. Enkelt sagt kan en si at en F-test vurder hvorvidt *noe som helst skjer*, altså har kovariatene noen som helst effekt på den avhengige variabelen (som bestemt av β-koeffisientene). Det er slik sett den enkleste hypotesetesten rent tolkningsmessig da spørsmålet er hvorvidt det finnes noe som helst påvirkning på den avhengige variabelen, og den mest åpenbare nærliggende hypotesen vil da være hvorvidt (for eksempel) bøker i hjemmet har en effekt på leseferdigheter.

$$H_o : \beta_1 = \beta_2 = ... = \beta_k = 0 \tag{2.21}$$

$$H_1 : \beta_j \neq 0 \text{ for iallfall en } j \;\; \epsilon\{1, ..., k\} \tag{2.22}$$

Null-hypotesen forkastes dersom kovariatene har en effekt på den forklarte variabelen. Formelt kan F-testen formuleres slik:

$$F = \frac{n-p}{k} - \frac{R^2}{1-R^2} \sim F_{k,n-p} \tag{2.23}$$

Der k er antall kovariater, p er $(k+1)$ (antall β parametere (koeffisienter) i dette tilfellet) og n er antall observasjoner. Ved verdier over *kritisk verdi* forkastes nullhypotesen. For F-testen kan de kritiske verdiene hvor null-hypotesen forkastes uttrykkes slik:

$$F > F_{k,n-p}(1-\alpha) \tag{2.24}$$

Som oftest finner vi derimot at vi legger til eller tar bort variabler fra regresjonen, slik at vi arbeider med to modeller hvor den ene er en undermodell av den andre.

Det gir derfor mening å forkaste en null-hypotese, om ekstra forklaringskraft som følge av å legge til flere variabler til den originale regresjonsligningen, er sterk nok. Vi kan måle denne "ekstra forklaringskraften" gjennom å betrakte en økning i forklart sum av kvadrater (SSE, som vist i 2.1.3). Siden total sum av kvadrater (SST) er den samme for begge modeller kan vi utlede at:

$$SST = SSE_u + SSR_u \quad \text{og} \quad SST = SSE_r + SSR_r \tag{2.25}$$

Om vi tar utgangspunkt i dette kan vi omformulere dette til:

$$SSE_u - SSE_r = (SST - SSR_u) - (SST - SSR_r) = SSR_r - SSR_u \tag{2.26}$$

Følgelig er det slik, at vi forkaster H_0 hvis den relative reduksjonen i SSR er tilstrekkelig stor, ved følgende test-observator (Thomas, 2005, s. 418)

$$F = \frac{SSR_r - SSR_u/h}{SSR_u/(n-k)} \tag{2.27}$$

Hvor n er antall observasjoner og k er antall kovariater, og gir sammen antall frihetsgrader for den utvidede modellen. h er forskjell mellom antall kovariater i den fulle modellen og den reduserte. Under H_0 har F en F-fordeling med $(h, n-k)$ frihetsgrader. Gir denne testen en verdi som er større enn en, gir dette to muligheter: enten er den utvidede modellen korrekt, eller så er den reduserte modellen korrekt, men tilfeldighet gjør at modellen passer bedre. Den kritiske verdien forteller oss hvor stor testobservatoren må være for at sannsynligheten for sistnevnte utfallsmulighet er tilstrekkelig lav, i praksis og som regel mindre enn 5%, det vil si $p < 0.05$

2.2 Skoleproduktfunksjonen og teoretiske innfallsvinkler til valg av variabler

2.2.1 Skoleproduktfunksjonen

Innen utdanningsøkonomi er det vanlig å bruke en alminnelig produktfunksjon ($Q = F(K, L)$) som utgangspunkt (Hanushek, 2020, s. 161), der elevprestasjoner uttrykkes som produktet av innsatsfaktorer i relevant skolesammenheng. Den kan uttrykkes slik:

$$Q_{it} = f(F_i^{(t)}, P_i^{(t)}, S_i^{(t)}, A_i) + v_{it} \tag{2.28}$$

Der Q_{it} er elevprestasjonene til student i ved tidspunkt t. $f(F_i^{(t)})$ uttrykker kumulativ familiekarakteristika frem til tid t. $P_i^{(t)}$ er kumulative medelevkarakteristika. $S_i^{(t)}$ er kumulativ skolekarakteristika, og A_i er iboende eller eksisterende ferdigheter. Til slutt følger et stokastisk ledd v_{it} (Hanushek, 2002, s. 2069). Formuleringen er altså bred, og en kan i utgangspunktet ta i bruk mange faktorer, selv de utenfor skolen. Funksjonen fungerer også over tid, som vil være meningsfylt med tanke på at læring ofte bygger på læring. Dette stiller derimot store krav til datamaterialet,

da en behøver historisk data for elever, som gjør det vanlig å sette opp modellen på value-added form (Hanushek, 2002, s. 2070; Bonesrønning, 2004, s. 16):

$$Q_{it} - Q_{it*} = f(F_i^{(t-t*)}, P_i^{(t-t*)}, S_i^{(t-t*)}) + e_{it} \tag{2.29}$$

Slik at iboende ferdigheter kumulativt opparbeidet faller bort og det blir ikke nødvendig med datamateriale på elevens karakteristikk før utdanningsforløpet. Det er derimot data tilgjengelig på tidlig leseevne i PIRLS-datasettet, så det vil være naturlig å ta den med, men de er likevel kun en av potensielt svært mange indikasjoner på elevens iboende evner. Sammenhengen mellom de ulike kategoriene av karakteristika og hvilke som utgjør vår modell redegjøres for under empirisk strategi.

2.2.2 Teoretisk utgangspunkt for problemstilling, hypoteser og variabler

Teoretisk følger vi Myrberg og Rosén (2009) samt Gustafsson mfl. (2011), i utarbeidelsen av vår problemstilling. De peker på at hjemmedynamikken er essensiell for å skape gode forutsetninger for barns leseevne. På linje med tidligere forskning (Adams, 1990; Coleman, 1968; Hanushek, 1989) isoleres sosioøkonomiske faktorer som den viktigste forklaringen på forskjeller i leseferdighet. Innenfor sosioøkonomiske forskjeller peker Myrberg og Rosén (2009) på foreldrenes utdanning som den viktigste faktoren. Utdanning har en tendens til å gli over i andre aspekter av familekarakteristika, og viser seg i flere former. En høyt utdannet husstand har ofte en høyere inntekt, samt flere bøker i hjemmet, som begge vil gi utslag på leseevne. For et stort flertall av barn med høyt utdannede foreldre, har flere mer enn 200 bøker i hjemmet (45 prosent), men hva med barn fra hjem der det ikke er rikelig tilgang på lesestoff? I denne oppgaven utforsker vi effekten av å ha *få* bøker i hjemmet.

Mangen (2018) peker på at elever som leser analogt scorer bedre på leseferdighetstester enn elever som leser digitalt. Samtidig er det en eksplisitt politisk ambisjon å øke PC tettheten i den norske skolen, for å skape det de kaller en "teknologisk skolesekk" (Kunnskapsdepartementet, 2017). Dette er ikke nødvendigvis hensiktsmessig for norske elevers leseferdigheter dersom denne PC-tettheten fører til mer tekst bearbeidet på skjerm. Kretzschmar mfl. (2013), sammenlignet i sin studie hvorvidt det var vanskeligere for gjennomsnittspersonen å prosessere tekst på skjerm kontra bok. Resultatet var at subjektene foretrakk å lese tekst på papir til tross for at det ikke nødvendigvis er noe lettere å lese analogt. Det viste seg også at noen hadde lettere for å lese på skjerm på grunn av bedre kontraster og bakbelysning. Dette gjenspeiles i studiet til Wiberg og Myrberg (2015) som argumenter for at det ikke nødvendigvis er tekst på skjerm i seg selv, men subjektets vaner og preferanser, som kan gi utslag på elevers leseferdigheter. Altså leser man best på den formen man er vant til å lese på.

Vi kan følgelig formulere følgende hypoteser for å formalisere forskningsspørsmålet:

H_0 : Bøker i hjemmet har ingen effekt på leseferdighet

H_1 : Færre bøker i hjemmet har en negativ effekt på leseferdighet

H_2 : Tilgang til PC i klasserommet har en positiv påvirkning på leseferdighet

For å fange effekten av sosioøkonomisk status på elevers leseferdigheter er det hensiktsmessig å fokusere på mengden bøker i hjemmet som uavhengig variabel. I vår utvidede modell inkluderer vi i tilegg indikatorer for inntekt, foreldrenes utdannelse, om norsk snakkes i hjemmet, hvorvidt eleven er født i utlandet, og elevens tidligere leseferdigheter. Dette mener vi er en egnet kombinasjon av variabler for å fange sosioøkonomisk status, og hvordan det påvirker elevers leseevne. Under følger en tabell over hvilke variabler vi inkluderer i modellen, og hvorfor vi har valgt dem:

Variabler	**Våre rasjonaler for å inkludere den i modellen**
books home	Dette er vår uavhengige variabel. Vi har valgt denne da den funker som en god indikator for de generelle sosioøkonomiske forholdene i hjemmet.
par edu	Denne variabelen inkluderes da den ofte samsvarer med et stort antall bøker i hjemmet. Foreldres nivå av utdanning sees på som en av de viktigste faktorene for en husholdnings kulturelle og sosiale kapital, som igjen gir et bedre utgangspunkt for leseferdigheter.
income	I likhet med books home og par edu, fanger income opp det generelle sosioøknomiske nivået i hjemmet. Denne variabelen er antagelig sterkt korrelert med nivå av utdannelse, og dermed antall bøker i hjemmet.
speak tesltang home	Hvorvidt språket eleven testes i snakkes i hjemmet vil ha en antatt effekt på deres evne til å lese og å forstå tekster på språket. Dette betyr ikke nødvendigvis at disse elevene er dårligere til å lese, men kan heller indikere at de ikke er like vant til å lese på et visst språk.
not born	Vi inkluderer en variabel på om eleven er født i landet de har testen i. I likhet med hvilke språk som snakkes i hjemmet vil elevens leseprestasjoner påvirkes av hvilke språk de er vant til å bearbeide tekst i.
early ability	Hvordan barn presterer tidlig i sin skolekarriere vil naturligvis påvirke deres leseevne i senere skoleforløp. Denne variabelen fanger opp sosioøkonomiske tilstander i hjemmet da den betegner hvor forberedt de var på skolegang før de begynte på skolen. Denne variabelen reflekterer ofte foreldres utdannelsesnivå.

3 Deskriptiv statistikk

I vårt forsøk på å undersøke effekten av bøker i hjemmet på leseferdigheter har vi benyttet oss av en rekke variabler fra PIRLS testen for Norge fra 2001. I dette kapittelet presenterer vi de benyttede variablene, og hvordan vi har bearbeidet dem for å skape en robust modell. Redegjørelsen for variablene vil bestå av deskriptiv statistikk, samt vårt rational for å inkludere dem i modellen.

Analysen er utarbeidet og basert på PIRLS datamaterialet for Norge fra 2001. PIRLS (Progress in International Reading Literacy Study) er en studie initiert av The International Association for 'the Evaluation of Educational Achievement (IEA)', med mål om å kartlegge leseferdighetene blant fjerdeklassinger på tvers av land. Det norske datasettet inneholder data fra 3459 respondenter tilhørende 136 forskjellige skoler (Solheim & Tønnesen, 2003).

3.1 Variabler

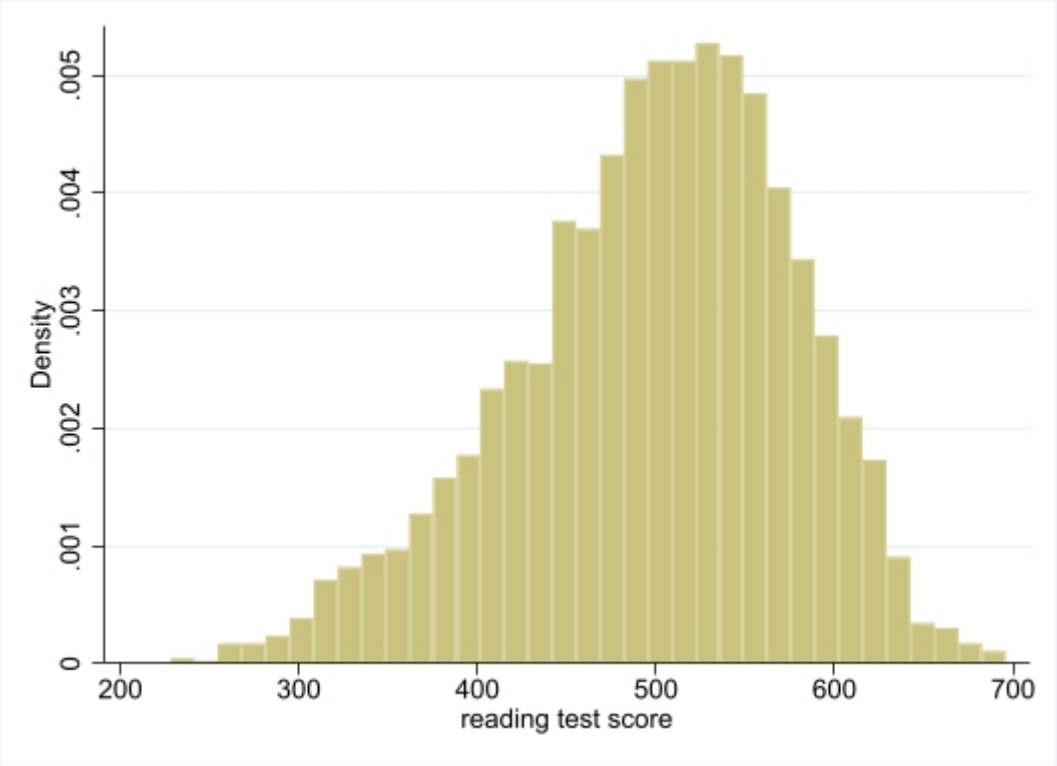
3.1.1 Avhengig variabel

PIRLS' leseferdighetsskala baserer seg på gjennomsnittlig leseferdigheter blant fjerdeklassinger for alle deltagende land, der resultatene sammenlignes for å skape et komparativt mål. Dette målet brukes deretter for å observere endringer i leseferdigheter over tid, og på tvers av land. Gjennomsnittlig leseferdighet på verdensbasis tillegges en verdi på 500. I 2001 så man en typisk variasjon på rundt 100 poeng. I Norge så vi i 2001 følgende resultater av PIRLS testen:

Variable	Obs	Mean	Std. Dev	Min	Max
Read	3459	498,2563	78,3662	228,0606	695,8717

Tabell 3.1.1: Deskriptiv statistikk for *read*

3459 elever deltok i den norske undersøkelsen. Gjennomsnittsscoren var 498, med et standardavvik på 78. Histogram over fordelingen:



3.1.2 Uavhengig variabel

Vår uavhengige variabel er books home, som er et mål på antall bøker i hjemmet. Tabell 3.1.2 viser deskriptiv statistikk for variabelen; tabell 3.1.3 viser fordelingen av respondenter i hver kategori:

Variable	Obs	Mean	Std. Dev	Min	Max
books home	3133	4,0328	1,0479	1	5

Tabell 3.1.2: Deskriptiv statistikk for *books home*

I vår analyse ser vi det hensiktsmessig å kode kategorivariabelen om til dummyvariabler. Vi koder books home til fem dummyvariabler i henhold følgende kategorier.

books home	Antall bøker	Freq	Percent
1	0 til 10 bøker	63	2,01
2	11 til 25 bøker	163	5,20
3	26 til 100 bøker	803	25,63
4	101 til 200 bøker	683	21,80
5	200 bøker eller mer	1421	45,36

Tabell 3.1.3: Fordeling av *books home*

Tabell 3.1.3 viser en overvekt av respondenter med mer enn 200+ bøker i hjemmet. Dermed velger vi å bruke denne kategorien som referansekategori i videre modellering og analyse.

3.1.3 Kontrollvariabler

Vi ønsker å ikke inkludere for mange kontrollvariabler for å la effekten av bøker i hjemmet på leseevne skinne. Analysen baserer seg på tre modeller, der den første ser på den isolerte effekten av antall bøker i hjemmet på leseferdigheter. Vi introduserer deretter i modell to, variablene *income*, *pc class*, *par uni* og *speak testlang home* for å undersøke flere mål innenfor familiekarakteristika. Deskriptiv statistikk for disse variablene er i tabell 3.1.4.

Variable	Obs	Mean	Std. Dev	Min	Max
Income	2994	4,0641	1,5507	1	6
pc class	3415	0,8497	0,3573	0	1
par uni	3098	0,5429	0,4982	0	1
speak testlang home	3389	1.1118	0,3573	1	3

Tabell 3.1.4: Deskriptiv stat. for *Income*, *pc class*, *par uni* og *speak testlang home*

Variabelen *income* måler foreldrenes inntekt i 2001 i USD (amerikanske dollar). I likhet med *books home* er *income* en kategorivariabel med fem kategorier. Tabell 3.1.4 viser deskriptiv statistikk for variabelen; tabell 3.1.5 viser antall respondenter i hver kategori:

income	Inntekt	Freq	Percent
1	LESS THAN $20,000	193	6,45
2	$20,000 - $29,999	370	12,36
3	$30,000 - $39,999	512	17,1
4	$40,000 - $49,999	625	20,88
5	$50,000 - $59,999	565	18,87
6	$60,000 OR MORE	729	24,35

Tabell 3.1.5: Fordeling av *income*

Vi ser at det er klar overvekt av mennesker som tjener over 40 000 USD i Norge. Det amerikanske målet bidrar til at flere nordmenn havner i de øverste kategoriene på grunn av at norske lønninger var gjennomsnittlig høyere enn amerikanske på tidspunktet undersøkelsen ble gjennomført. Derfor anser vi det hensiktsmessig å snevre denne inn i én dummyvariabel som måler om foreldrene tjener mer eller mindre enn 49.000 USD. Innsnevringen kan forklares ved at de aller fleste nordmenn tjener mer enn 49.000 2001-dollar. Dermed fanger vi opp effekten av å komme fra et hjem med gjennomsnittlig eller lav inntekt.

Variabelen *pc class* er en dummyvariabel og gir innsikt i om elevene har tilgang til PC i klasserommet. 84 prosent av de spurte elevene har svart positivt.

Variabelen *par uni* er dummykodet fra variabelen *par edu*. Dummyen fanger opp om barnets foreldre har høyere utdanning eller ikke. I likhet med inntektsvariabelen ser vi på dette som hensiktsmessig da en stor overvekt av respondentene hadde høyere utdanning. Tabell 3.1.4 viser at hele 54 prosent av respondentene tilhører

denne kategorien.

Variabelen *speak testlang home* er også en kategorivariabel. Respondentene har svart på om de snakker testspråket hjemme i sin egen husholdning. Tabell 3.1.6 viser fordelingen av respondentenes svar.

Speak testlang home	Nivå	Freq	Percent
1	Alltid	3058	90,23
2	Noen ganger	283	8,35
3	Aldri	48	1,42

Tabell 3.1.6: Fordeling *speak testlang home*

Svært få snakker aldri eller noen ganger norsk hjemme. Det er nærliggende å tro at respondenter som aldri snakker norsk hjemme, eller kun gjør det noen ganger, scorer lavere enn de som alltid snakker norsk hjemme.

I modell 3 inkluderer vi andre variabler som ikke nødvendigvis faller innenfor familiekarakteristika, men som kan bidra til en bedre modell. Vi har inkludert følgende variabler *early1*, *early2*, *early3*, *early4* og *not born*. Variablene *early1-4* er *early ability* dikotomisert. Inklusjonen er naturlig, da vi ønsker å forsikre oss om at effekten av antall bøker i hjemmet på elever leseferdigheter ikke er optimistisk. Tabell 3.1.7 viser deskriptiv statistikk for disse variablene.

Variable	Obs	Mean	Std. Dev	Min	Max
early1	3111	0,1340	0,3407	0	1
early2	3111	0,3085	0,4619	0	1
early3	3111	0,3612	0,4804	0	1
early4	3111	0,1960	0,3970	0	1
not born	3355	0,0909	0,2875	0	1

Tabell 3.1.7: Deskriptiv statistikk *early1*, *early2*, *early3*, *early4* og *not born*

Variabelen *early ability* har antakeligvis stor effekt på leseevnen. Denne variabelen måler hvorvidt subjektet hadde evnen til å lese i tidlig alder. For å gi en bedre illustrasjon av effekten av å kunne lese i tidlig alder på leseferdighetene til en fjerdeklassing, har vi valgt å kode den om til dummier. Tabell 3.1.8 viser fordelingen av respondentene i de forskjellige kategoriene.

Early Ability	Nivå	Freq	Percent
1	Not at all	417	13,4
2	Not very well	960	30,86
3	Moderately well	1124	36,13
4	Very well	610	19,61

Tabell 3.1.8: Fordeling *early ability*

Som vi ser av tabellen er det spredning i nivået på leseferdigheten i tidlig alder.

Likevel er det verdt å merke seg at overvekten av respondentene kunne lese godt eller veldig godt.

Not born indikerer hvorvidt vedkommende er født i Norge eller ikke. Kun 9 prosent av respondentene er ikke født i Norge. Denne inkluderer vi da det er nærliggende å tro at en elev som har hatt mindre tid med språket vil ha større utfordringer med å lese og forstå det.

I tabell 8.0.1 i appendix presenter vi ytterligere kontrollvariabler. Her er alle kontrollvariabler som blir benyttet i robusthetstesten gruppert i kategorier og forklart. Kategoriene er henholdsvis familiekarakteristika, medelevkarakteristika og skolekarakteristika.

Oppgaven går bredt ut for å teste robustheten til kjernemodellen. Dette anser vi som svært viktig for å kunne stole på de resultatene vi finner i analysen senere i oppgaven. Robusthetstesten blir presentert i punkt 5.1.

3.2 Korrelasjonsmatrise

Variabler	**read**
read	1,0000
Dbooks1 10	-0,1033
Dbooks11 25	-0.1358
Dbooks26 100	-0.1740
Dbooks101 200	-0.0084
Dbooks200plus	0.2452
incomehigh2	0.2022
pc class	-0.0309
par uni	0.2914
speak testlang home	-0.1734
not born	-0.1119
early1	-0.2100
early2	-0.1973
early3	0.0671
early4	0.3270

Tabell 3.2.1: Korrelasjonsmatrise for variabler

Den sterkeste positive korrelasjonen finner vi mellom *read* og *early4*, fulgt av *par uni*. Resultatet er intuitivt, da det er nærliggende å tro at dersom man leste på et godt nivå tidlig, vil man også score godt i dag. I tillegg er det naturlig at dersom man har foresatte med utdannelse fra universitet eller høyskole, vil man score høyere. Det største negative korreleringen får man dersom man ikke alltid snakker norsk hjemme. Overraskende ser vi at alle nivåer av bøker i hjemmet,

bortsett fra *Dbooks200plus*, har en negativ effekt på *read*, isolert sett. Likevel er det tydelig at den laveste negative korrelasjonen, er når man har mellom 101 og 200 bøker i hjemmet. Vi ser videre at dersom man har over 200 bøker vil det være en positiv korrelasjon. Dette gir en indikasjon på at flere bøker bidrar til bedre leseferdigheter.

4 Strategi

I denne delen av oppgaven knyttes innledende generell teori sammen med det konkrete datamaterialet og variablene. Deretter settes dette i sammenheng med hypotesene som beskrevet i innledningen. Altså forklarer dette segmentet hvordan teorien og problemstillingen sees i sammenheng.

4.1 Økonometrisk grunnmodell

Som vist i 2.2.1 er den generelle skoleproduktfunksjonen utgangspunktet for modellen vår, hvor elevprestasjonen Q forståes som en funksjon f av ulike karakteristika. Dermed er sammenhengen i enklest mulig forstand å forstå som $L = f(F)$ hvor L uttrykker leseferdighet og bøker i hjemmet inngår i familiekarakterika F. Betrakter vi kun avhengig og uavhengig variabel for vår modell, bøker i hjemmet og leseferdighet som elevprestasjon, kan dette da uttrykkes som:

$$L_i = f(F_i) = \beta_0 + \beta_a B_i \qquad a = 1, 2, 3, 4 \tag{4.1}$$

$$L_i = f(F_i) = \beta_0 + \beta_1 B_i + \beta_2 B_i + \beta_3 B_i + \beta_4 B_i \tag{4.2}$$

Hvor L_i uttrykker leseferdighet for elev i og β_0 er konstantleddet. $a = 1, 2, 3, 4$ uttrykker de diktomiserte koeffesientene β_1, β_2, β_3 og β_4 for henholdsvis 1-10 bøker i hjemmet, 11-25 bøker i hjemmet, 26-100 bøker i hjemmet og 101-200 bøker i hjemmet. Ettersom 200+ bøker i hjemmet er vanlig, er dette som nevnt referansekategori, og finner elev i seg i denne kategorien, er leseferdighet uttrykt som konstantleddet som utgangspunkt. Ligning (4.2) skriver dette helt ut for ordens skyld, men for videre utledning er dette vist som i (4.1) og andre dikotome variabler er slått sammen på samme måte. Vi introduserer gradvis flere variabler i hver modell, først og i hovedsak familiekarakteristika slik argumentert for, samt noe elevkarakteristika, og modell 3 kan dermed uttrykkes slik:

$$L_i = \beta_0 + \beta_a B_i + \beta_5 P_i + \beta_6 U_i + \beta_7 I_i + \beta_8 S_i + \beta_9 V_i + \beta_b E_i \tag{4.3}$$

Variablene er som overnevnt, men i tillegg introduserer vi da: P for tilgang til PC i skolen; U som er hvorvidt foreldre har universitetsutdannelse; I som tar verdi 1 om inntekt er 50 000$ eller mer; S angir hvorvidt eleven snakker språket testen gis på hjemme; V hvorvidt eleven er født i landet; og E angir tidlig leseevne, hvor $b = 10, 11, 12$ og angir henholdsvis hvorvidt tidlig leseevne er ikke-eksisterende, dårlig eller svært godt, med nokså god som referansekategori. Alle for elev i.

4.2 Formulering av hypotese

Som forklart i innledning og teori, har vi valgt å formulere en null-hypotese, en hovedhypotese og en alternativ hypotese, disse testes etter regresjonsanalysen:

H_0 : Bøker i hjemmet har ingen effekt på leseferdighet

H_1 : Færre bøker i hjemmet har en negativ effekt på leseferdighet

H_2 : Tilgang på PC i klasserommet har en positivt påvirkning på leseferdighet

5 Data

I denne delen av oppgaven presenterer vi først en robusthetstest og deretter en endelig modell som skal benyttes i analysen.

5.1 Robusthettest

For å sjekke om effekten av familiekarakteristika på *read* er robust foretar vi først en regresjon der vi kontrollerer for medelevkarakteristika i modell 2, og for skolekarakteristika i modell 3. Vi observerer at effekten av bøker i hjemmet på *read* er tilnærmet konstant gjennom modellene. Det samme ser vi for de andre familiekarakteristikavariablene, som også holder seg tilnærmet like gjennom alle modellene. Variablene i hovedmodellen forblir signifikante på 0.01 nivå. Vi ser i tilegg en minimal økning R^2 på tross av inklusjonen av flere variabler. Tilsammen indikerer dette at modellen vår er robust.

Variables	(1) **read**	(2) **read**	(3) **read**
Dbooks1_10	-40.51**	-40.26**	-43.72**
	(9.692)	(10.46)	(12.42)
Dbooks11_25	-33.58**	-32.06**	-35.41**
	(6.265)	(6.608)	(8.072)
Dbooks26_100	-19.90**	-20.68**	-18.88**
	(3.425)	(3.697)	(4.308)
Dbooks101_200	-10.19**	-11.76**	-8.685*
	(3.337)	(3.573)	(4.096)
par_uni	27.66**	26.60**	26.37**
	(2.915)	(3.133)	(3.675)
par_not_born	-13.56*	-18.38*	-19.26*
	(6.772)	(7.234)	(8.209)
early1	-46.45**	-46.80**	-43.73**
	(4.049)	(4.334)	(5.018)
early2	-28.92**	-29.80**	-28.19**
	(3.084)	(3.332)	(3.842)
early4	38.72**	37.65**	38.88**
	(3.518)	(3.773)	(4.363)
incomehigh2	10.95**	12.05**	11.35**
	(2.784)	(3.029)	(3.496)
speak_testlang_sometimes	-20.80**	-21.08**	-18.89**
	(5.278)	(5.661)	(6.609)
speak_testlang_never	-69.50**	-68.59**	-73.21**
	(12.94)	(13.50)	(14.53)
parent_ujob	-3.299	-2.230	-3.033
	(6.924)	(7.492)	(8.915)
pct_abroad26_50		10.30*	3.828
		(4.541)	(5.606)
pct_abroad50plus		10.91	-2.238
		(13.05)	(15.57)
pct_disadv26_50		-2.757	-4.504
		(4.231)	(5.095)
pct_disadv50plus		-6.332	5.995
		(10.79)	(13.57)
pc_class			-13.21**
			(4.491)
teacher_cert			9.189
			(11.47)
teacher_exp			-0.235
			(0.145)
clsize			-0.415
			(0.372)
school_location0_3000			5.274
			(4.627)
school_location100001_500000			15.96**
			(4.495)
school_location500000plus			8.825
			(6.588)
Constant	504.4**	504.9**	514.2**
	(3.463)	(3.816)	(15.97)
Observations	2,728	2,379	1,779
R-squared	0.274	0.279	0.288

Standard errors in parentheses
** p<0.01, * p<0.05, * p<0.10

Tabell 5.1.1: Modell testet for robusthet

Vi ser at det i hovedsak er familiekarakteristika som er utslagsgivende for elevers leseferdigheter, derfor kommer vi i analysen til å basere oss på den mindre modellen presentert i 5.2.1.

5.2 Endelig modell

VARIABLES	(1) **read**	(2) **read**	(3) **read**
Dbooks1_10	-79.00**	-47.20**	-41.74**
	(9.584)	(10.21)	(9.473)
Dbooks11_25	-67.39**	-39.67**	-32.19**
	(6.248)	(6.535)	(6.191)
Dbooks26_100	-46.02**	-26.08**	-22.74**
	(3.312)	(3.620)	(3.368)
Dbooks101_200	-24.02**	-13.56**	-10.07**
	(3.494)	(3.585)	(3.335)
pc_class	-4.194	-3.230	-2.193
	(3.763)	(3.792)	(3.534)
par_uni		29.34**	28.03**
		(3.109)	(2.892)
incomehigh2		11.40**	11.04**
		(2.978)	(2.769)
speak_testlang_home		-32.90**	-29.56**
		(4.126)	(3.970)
not_born			-15.16**
			(4.757)
early1			-49.65**
			(4.012)
early2			-29.64**
			(3.066)
early4			37.10**
			(3.510)
Constant	527.3**	533.1**	537.3**
	(3.741)	(6.423)	(6.302)
Observations	3,089	2,863	2,786
R-squared	0.090	0.147	0.280

Standard errors in parentheses
** p<0.01, * p<0.05, * p<0.10

Tabell 5.2.1: Endelig modell

6 Resultat

I denne delen vil vi analysere og tolke regresjonsresultatene for de tre ulike modellene. Deretter gjennomfører vi hypotesetest for å gjøre relevante slutninger for problemstillingen.

6.1 Regresjonsanalyse

Tabell 5.2.1 viser våre 3 grunnmodeller og regresjonen av disse.

I den første modellen utfører vi en regresjon mellom vår avhengige og uavhengige variabel, henholdsvis *read* og *Dbooks1 10, Dbooks 11 25, Dbooks26 100 og Dbooks 101 200*. I tråd med våre forventinger viser modelen at effekten av bøker i hjemmet på elevers leseferdigheter både er stor, og signifikant. Elever scorer progressivt lavere på leseferdighet ettersom bøker i hjemmet minker. Dersom Dbooks er dummykodet vil tolkingen av variablene være at alle elever vil høre til én kategori, som trukket fra konstantleddet. Dersom man kun har mellom 1 og 10 bøker i hjemmet vil man da i gjennomsnitt score 444.57 på leseferdighettesten. Grunnen til at effekten av kategoriene er negativ, er at vi bruker hjem med 200plus bøker som referansekategori. Denne ble valgt som referanse da et klart overtall av elever (45 prosent) kommer fra et hjem med 200plus bøker. Det vil si at modellen viser effekten av å ikke tilhøre referansekategorien. Vi merker oss at alle variabler, bortsett fra *pc class* er signifikante på 0.01 nivå og at R^2 er 0.090. Effekten av få bøker i hjemmet vil her være *veldig* optimistisk da vi ikke kontrollerer for andre variabler. Dette betyr at effekten av andre variabler vil tilegges dbooks i denne modellen. Vi ser av modellen at *pc class* er negativ, men insignifikant. Som vi ser i videre utvidelser vil effekten av bøker i hjemmet fortsatt være svært signifikant, dog ha en litt mindre estimert effekt på read.

I modell 2 introduserer vi flere variabler for å skape et større begrep for familiekarakteristika. Dette inkluderer foreldres inntekt, utdannelse og hvor ofte man snakker norsk hjemme. Som ventet ser vi en positiv effekt av variablene *par uni* og *incomehigh2*. Barn som kommer fra hjem der de ikke snakker norsk kommer dårligere ut i leseferdigheter. Disse resultatene er i tråd med vår hypotese om at familiekarakteristika har mye å si for elevenes leseferdighet. Vi observerer at den estimerte effekten av bøker i hjemmet på *read* minker, da vi inkluderer flere variabler som fanger opp effekt av familiekarakterestika, som tideligere var tillagt bøker i hjemmet. R^2 i modell 2 er 0.134, altså høyere enn i modell 1. Dette er intuitivt da man ved inklusjon av flere variabler vil forklare en større andel av variansen i *read*. Vi merker oss at alle variabler, untatt *pc class*, er signifikante på 0.01 nivå.

I modell 3 utvider vi modellen med *not born* og *early ability*. Disse variablene omhandler i større grad eleven som person. Vi ser at den estimerte effekten av å ha gode leseferdigheter på starten av skoleløpet er svært positiv. Tabell 3.1.7 viser fordelingen av *early ability*. Modellen viser en klar sammenheng mellom tidlig og nåværende leseferdighet. I og med at variabelen er dummykodet vil hver enkelt elev se en effekt av tidlig leseevne i henhold til kategorien de tilhører. Grunnen til at effekten av *early1* og *early2* er negativ, og *early4* er positiv, er fordi vi bruker *early3* som referansekategori, da de aller fleste befinner seg i denne kategorien. Vi velger i denne oppgaven å se på dette som en indirekte indikator på familiekarakteristika

da det å være flink til å lese i ung alder kan tillegges hjemmets og foreldrenes sosioøkonomiske kapital. Foreldres utdannelse og inntekt har fortsatt en stor effekt på leseferdigheter. Det samme gjelder for effekten av bøker der vi ser en liten reduksjon i effekt på leseferdigheter ved å inkludere flere variabler. Den reduserte effekten er liten i forhold til økningen i R^2 som betyr at den negative effekten er robust. Alle variablene i modellen er fremdeles signifikante, med unntak av *pc class*.

Vi har valgt å inkludere *pc class* i alle regresjonsmodellene, til tross for at variabelen ikke er signifikant. Dette ble gjort for å adressere hypotese 2. Vi undersøker om tilgang på PC i klasserommet har en positiv innvirkning på fjerdeklassingers leseferdigheter. Av modellene fremkommer det at det har negativ effekt på leseferdighetene dersom elever har tilgang på PC i klasserommet, vi kan ikke konkludere med at så er tilfellet, fordi variabelen ikke er signifikant. En interessant observasjon er at variabelen *pc class* er signifikant i robusthetstesten, etter inklusjon av både familiekarakteristika, medelevkarakteristika og skolekarakteristika.

6.2 Hypotesetest

Hypotesene er som nevnt:

H_0 : Bøker i hjemmet har ingen effekt på leseferdighet

H_1 : Færre bøker i hjemmet har en negativ effekt på leseferdighet

H_2 : Tilgang på PC i klasserommet har en positivt påvirkning på leseferdighet

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 \qquad = 0 \tag{6.1}$$

$$H_1 : \beta_1, \beta_2, \beta_3, \beta_4 \qquad \text{minst en av de} \qquad \neq 0 \tag{6.2}$$

Følgelig kan vi gjøre en F-test overfor null-hypotesen som redegjort for i teori-delen 2.1.5.1.

$$F = \frac{SSR_r - SSR_u/h}{SSR_u/(n-k)} \tag{6.3}$$

$$F = \frac{(12365432.9 - 12044364.3)/4}{(12044364.3)/2773} = 18.48007926 \tag{6.4}$$

Siden dette tallet er mye større enn 1, gir dette som vist i teori-delen angående hypotesetesting, to muligheter: den mer komplekse modellen som inneholder bøker i hjemmet passer best, eller den enklere modellen passer best, men tilfeldighet gjør at modellen passer bedre. Den kritiske verdien forteller oss hvor sjelden sistnevnte mulighet er (antar 0.05 som signifikansnivå): H_0 forkastes dersom den utregnede verdien $F = 18.48007926$ er større enn den teoretiske kvantilen $F_{h,n-k}(0.95) = F_{4,2773}(0.95)$. Den teoretiske kvantilen vil i dette tilfellet være 2.37, som lest fra

tabell eller STATA. Intuisjonen her vil være at selv om vår F-verdi er større enn 1, er spørsmålet om F-verdien er *signifikant* større enn 1. Her vil alt større enn 2.37 være signifikant og vi må forkaste vår null-hypotese i favør for hovedhypotesen H_1 som sier at færre bøker i hjemmer *har en negativ effekt på leseferdighet.* Som vist i den endelige modellen og forklart i regresjonsanalysen er variabelen pc_class ikke signifikant i utgangspunktet, og vi kan ikke si noe om hvorvidt vår alternativ hypotese H_2 kan beholdes eller ikke.

7 Diskusjon og Konklusjon

Utgangspunktet for oppgaven er å undersøke om færre antall bøker i hjemmet har en negativ effekt på elevens leseferdighet. Analysen og hypotesetesten viser tydelig en negativ korrelasjon mellom færre bøker i hjemmet og leseferdigheter. Modellen er statistisk signifikant og robust.

Det er dog betimelig å spørre seg om effekten av antall bøker i hjemmet på leseferdigheter er isolert. Det er ofte sånn at ressurssterke foreldre, de med god inntekt og høyere utdannelse, besitter flere bøker. Årsakssammenhengen her er ikke nødvendigvis at flere bøker fører til bedre leseevne, men heller at flere bøker er en karakteristikk for hjem med høy sosioøkonomisk kapasitet, og at slike hjem skaper barn som er flinke til å lese. Sagt med andre ord, kan man putte så mange bøker som helst i et hjem, uten at noen nødvendigvis kommer til å lese dem. Med det datasettet vi har tilgjengelig for denne oppgaven, føler vi at vi ikke har mulighet for å bygge en mer solid teori som på en tilstrekkelig måte fanger dette aspektet, da årsaksforklaringen til flere bøker i hjemmet ha sitt opphav i svært mange faktorer utenfor modellen og datasettet. Det er dog nærliggende å tenke at flere bøker i hjemmet vil samsvare med et mer ressurssterkt hjem, som vil gi utslag på leseferdigheter. Dette viser seg tilsynelatende i vår modell da *books home*, *par uni* og *incomehigh2* har store positive virkninger på *read*.

Foreldre som besitter slike typer karakteristikker er i mange tilfeller også i større grad i stand til å følge opp barna sin læring fra ung alder som viser seg i den store effekten av *early read*. Det blir tydelig at elever med et dårlig utgangspunkt i sin skolekarriere ofte får negative konsekvenser for senere resultater. Man kan spekulere i hvorvidt de som leste dårlig i ung alder gjør det nettopp fordi de ikke hadde tilgang på et tilstrekkelig antall bøker i hjemmet, men vi tenker det heller er et symptom på dårligere forutsetninger fra hjemmet. På den andre siden ser vi intuitivt nok, en stor positiv effekt av å kunne lese meget godt i tidlig alder; det er den største enkelt-faktoren til positiv score i modellen. Dette gir mening da læring er en kumulativ prosess så man vil alltid dra nytte av et godt utgangspunkt å bygge videre på.

Bøker i hjemmet som en variabel mangler nyanse. Den sier ingenting om det er

bøkene tilpasset barn eller voksne, eller hvilke språk de er på. Man kan jo tenke seg at barn som bare har tilgang på bøker rettet mot voksne, vil ha det vanskeligere for å lære seg å lese. Hvorvidt det er noe hold i dette er usikkert da det er mye som tyder på at det ikke er bøkene i seg selv som er utslagsgivende for leseferdighet, men hjemmemiljøet i huset de oppbevares i.

Når det gjelder tilgang på PC i klasserommet finner ikke vi grunnlag i vår modell for kommentere på dens effekt på leseferdigheter. I den utvidete modellen der alle karakteristikaene var inkludert så vi at tilgang til PC kan ha en negativ effekt på leseferdigheter og der viste den seg signifikant. Vi føler dog at vi ikke har et tilstrekkelig datagrunnlag for å dra noen konklusjoner om datamaskiners effekt i skolehverdagen. Vi legger dermed H_2 på is da vi hverken kan bekrefte eller avkrefte hypotesen.

Referanser

Adams, M. J. (1990). Beginning to read : thinking and learning about print. *Cambridge, Mass*, MIT Press.

Bonesrønning, H. (2004). Utforming av utdanningspolitikken - hva kan økonome bidra med? *Økonomisk forum*, *58*(3).

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

Gustafsson, J.-E., Yang, K. H. & Rosén, M. (2011). Chapter 4: Effects of Home Background on Student Achievement in Reading, Mathematics, and Science at the Fourth Grade. I S. Bradley & C. Green (Red.), *TIMMS and PIRLS 2011: Relationships Among Reading, Mathematics and Science Achievement at the Fourth Grade - Implications for Early Learning* (s. 181–187). Boston College Chestnut Hill, TIMSS & PIRLS International Study Center.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational researcher*, *18*(4), 45–62.

Hanushek, E. A. (2002). Chapter 30: Publicly provided education. I A. Auerbach & M. Feldstein (Red.), *Handbook of Public Economics* (4. utg., s. 2045–2414). Stanford University & NBER.

Hanushek, E. A. (2020). Chapter 13: Education Functions. I S. Bradley & C. Green (Red.), *Economics of Education* (2. utg., s. 161–170). London, Academic Press.

Kretzschmar, F., Pleimling, D., Hosemann, J., Fussel, S., Bornkessel-Schlesewsky, I. & Schlesewsky, M. (2013). Subjective Impressions Do Not Mirror Online Reading Effort: Concurrent EEG-Eyetracking Evidence from the Reading of Books and Digital Media. *PLoS ONE*, *8*(2).

Kunnskapsdepartementet. (2017). *Framtid, fornyelse og digitalisering - Digitaliseringsstrategi for grunnopplæringen 2017-2021*. Oslo, Kunnskapsdepartementet.

Mangen, A. (2018). Modes of writing in a digital age: The good, the bad and the unknown. *First Monday*, *23*(10).

Myrberg, E. & Rosén, M. (2009). Direct and indirect effects of parents' education on reading achievement among third graders in Sweden. *British Journal of Educational Psychology*, *79*(4), 695–711.

Solheim, G. R. & Tønnesen, F. E. (2003). *PIRLS: En norsk kortversjon av den internasjonale rapporten om 10-åringers lesekunnskaper*. Universitetet i Stavanger, Senter for leseforskning.

Thomas, R. (2005). *Using statistics in economics*. London, McGraw-Hill Ed.

Turmo, A. & Lie, S. (2004). *Hva kjennetegner norske skoler som skårer høyt i PISA 2000?* (Bd. 1/2004). Oslo, UiO/ILS.

Wiberg, N. & Myrberg, C. (2015). Screen vs. paper: what is the difference for reading and learning? *Insights*, *28*(2), 49–54.

8 Appendix

Karakteristika *Familie:*	**Variabler**	**Beskrivelse**
	Dbooks1 10	Dummy, tar verdi 1 dersom respondenten har 1-10 bøker hjemme, 0 ellers
	Dbooks11 25	Dummy, tar verdi 1 dersom respondenten har 11-25 bøker hjemme, 0 ellers
	Dbooks26 100	Dummy, tar verdi 1 dersom respondenten har 26-100 bøker hjemme, 0 ellers
	Dbooks101 200	Dummy, tar verdi 1 dersom respondenten har 101-200 bøker hjemme, 0 ellers
	par uni	Dummy, tar verdi 1 dersom foresatte har utdanning fra universitet eller høyskole, 0 ellers
	par not born	Dummy, tar verdi 1 dersom foresatte er født i utlandet, 0 ellers
	incomehigh2	Dummy, tar verdi 1 dersom foresattes inntekt er $50 000 eller mer, 0 ellers
	speak testlang sometimes	Dummy, tar verdi 1 dersom eleven snakker norsk hjemme noen ganger, 0 ellers
	speak testlang never	Dummy, tar verdi 1 dersom eleven aldri snakker norsk hjemme, 0 ellers
	parent ujob	Dummy, tar verdi 1 dersom foresatte er uten jobb, 0 ellers
Medelev:		
	pct abroad26 50	Dummy, tar verdi 1 dersom andel innvandrere ved skolen er 26-50%, 0 ellers
	pct abroad50plus	Dummy, tar verdi 1 dersom andel innvandrere ved skolen er over 50%, 0 ellers
	pct disadv26 50	Dummy, tar verdi 1 dersom andel elever fra fattige husholdninger ved skolen er 26-50%, 0 ellers
	pct disadv50plus	Dummy, tar verdi 1 dersom andel elever fra fattige husholdninger ved skolen er over 50%, 0 ellers
Skole:		
	pc class	Dummy, tar verdi 1 dersom det er pc tilgjengelig i klasserommet, 0 ellers
	teacher exp	Kontinuerlig, indikerer antall år læreren har jobbet som lærer
	clsize	Kontinuerlig, indikerer hvor mange elever det er i klassen
	school location 0-3000	Dummy, tar verdi 1 dersom det bor mellom 3000 og 100000 mennesker i skolens område, 0 ellers
	school location 100001-500000	Dummy, tar verdi 1 dersom det bor mellom 100001 og 500000 mennesker i skolens område, 0 ellers
	school location 500000plus	Dummy, tar verdi 1 dersom det bor over 500000 mennesker i skolens område, 0 ellers

Tabell 8.0.1: Forklaring av variabler