Bacheloroppgave

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

Tobias Fjellstad Back
Tobias Gamrath
Stian Topland
Øystein Sand

Komparativ analyse av prestasjonsforskjeller mellom innvandrerelever og øvrige elever i Norge og Tyskland

Bacheloroppgave i samfunnsøkonomi
Veileder: Bjarne Strøm

Mai 2020

NTNU
Kunnskap for en bedre verden

Tobias Fjellstad Back
Tobias Gamrath
Stian Topland
Øystein Sand

Komparativ analyse av prestasjonsforskjeller mellom innvandrerelever og øvrige elever i Norge og Tyskland

Bacheloroppgave i samfunnsøkonomi
Veileder: Bjarne Strøm
Mai 2020

Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
Institutt for samfunnsøkonomi

NTNU

Kunnskap for en bedre verden

NTNU
Kunnskap for en bedre verden

SØK2901 - Bacheloroppgave i samfunnsøkonomi

Komparativ analyse av prestasjonsforskjeller mellom innvandrerelever og øvrige elever i Norge og Tyskland

FORFATTERE

Stian Topland

Tobias Fjellstad Back

Tobias Gamrath

Øystein Sand

DATO

14.05.2020

ANTALL SIDER

30 sider hvorav 5 sider er vedlegg

SAMMENDRAG

I løpet av de første årene på grunnskolen skal elevene gå fra å lære å lese, til å lese for å lære. Å mestre dette er svært viktig for både resten av skolegangen og arbeidslivet senere. I denne analysen har vi derfor undersøkt leseferdighetene til innvandrere og ikke-innvandrere i Norge og Tyskland, med utgangspunkt i datamateriale fra PIRLS-undersøkelsen i 2001. Hensikten har vært å se om det eksisterer et prestasjonsgap mellom innvandrere og ikke-innvandrere i Norge og Tyskland, om det er ulikheter mellom landene, og hva som eventuelt påvirker resultatene.

I vår analyse observerer vi et prestasjonsgap mellom innvandrere og ikke-innvandrere. Vi finner ut at det ikke er en signifikant forskjell i gapet mellom Norge og Tyskland. Det vil si at differansen i leseferdigheter mellom innvandrere og ikke-innvandrere ikke er signifikant ulik mellom Norge og Tyskland. I begge land presterer jenter bedre enn gutter, men gapet er større i Norge enn i Tyskland. Videre har vi sett på utvalgte sosiodemografiske variabler hvor flere har vist seg å ha en signifikant effekt på leseferdighetene.

Innholdsfortegnelse

1. Innledning

1.1.Motivasjon

Gode leseferdigheter er viktig for å kunne tilegne seg kunnskap. Dersom elever sliter med lesing, vil de kunne få utfordringer med å prestere godt på både skolen og i arbeidslivet senere. I vår oppgave ønsker vi derfor å undersøke om leseferdighetene til innvandrerelever skiller seg fra øvrige elever, ved å sammenlikne Norge og Tyskland. Motivasjonen er at vi ønsker å se om det er noen signifikante forskjeller mellom to land med relativt likt velstandsnivå (BNP per innbygger), og som førte forholdsvis lik innvandringspolitikk på tidspunktet for undersøkelsen. Både Tyskland og Norge førte på 90-tallet en ganske restriktiv innvandringspolitikk, men har siden rundt år 2000 ført en mindre restriktiv innvandringspolitikk for å skape økonomisk vekst. (BPB, 2018), (SSB, 2013, s.53) & (Migrationpolicy, 2016)

1.2.Problemstilling

Oppgaven har som formål å besvare følgende problemstilling:
"Eksisterer det et prestasjonsgap mellom innvandrere og øvrige elever i Norge og Tyskland, og i hvilken grad påvirkes dette i så fall av sosiodemografiske forskjeller?"
I denne oppgaven skal vi undersøke hvilke forhold, som kan ha betydning for forskjeller mellom innvandrerelever og øvrige elever i Norge og Tyskland. Ved å analysere ulike sosiodemografiske variabler som kjønn, innvandrerstatus og husholdningers inntekt, ønsker vi å se om et eventuelt prestasjonsgap kan forklares ved hjelp av de variablene vi bruker i analysen.

Datamaterialet vi bruker i denne oppgaven er fra den internasjonale leseundersøkelsen PIRLS i 2001. Resultater fra nasjonale prøver i 2019 viser at elever med innvandrerbakgrunn jevnt over presterer svakere enn øvrige elever. Hvordan elever presterer på skolen i tidlig alder er ofte en pekepinn på hvordan det vil gå senere i livet for den enkelte eleven. Derfor har det også vært mye forskning rundt hva som påvirker elevprestasjoner tidlig i skoleløpet. Dette er blant annet hva PIRLS-undersøkelsen kartlegger. (SSB, 2019)

Innledningsvis ønsker vi å undersøke hvilke sosiodemografiske variabler som påvirker leseferdighetene. Dernest vil vi finne ut om det er forskjeller i variablenes påvirkningskraft i Norge og Tyskland. Tyskland har en del fellestrekk med Norge når det kommer til skolesystemet, og er

dermed sammenlignbart. De aller fleste tyske skoler er offentlige slik som i Norge og elevene begynner på grunnskolen året de fyller seks år. (Pedersen, 2018)

2. Teoretisk rammeverk og tidligere litteratur

2.1. Forskning på skoleprestasjoner for innvandrere

I dette kapittelet skal vi presentere teori og empiri knyttet til prestasjonsforskjeller mellom innvandrerelever og øvrige elever. Vi vil presentere forskningsresultater som omhandler årsakene til prestasjonsgapet og hva som påvirker elevprestasjonene.

De siste 50 årene har innvandringen økt betraktelig i Norge, og bare de siste 15 årene har andelen barn og unge med innvandringsbakgrunn mer enn fordoblet seg. En av seks elever kommer fra en familie hvor begge foreldrene har innvandret til landet. Tall fra 2016 viser også at disse elevene jevnt over presterer svakere på nasjonale prøver og andre kartleggingsprøver enn ikke-innvandrere. I tillegg har de lavere karaktersnitt på ungdomsskolen og høyere frafall på videregående skole, spesielt blant gutter. (UDIR, 2016, s. 2)

Rapporten "*Slik leser 10-åringer i Norge*" fra 2003 kartlegger resultatene fra PIRLS-undersøkelsen som ble gjennomført i 2001. Et relevant funn var knyttet til språket som ble snakket i hjemmet. 10-åringene ble de delt inn i to grupper; de som alltid snakker norsk hjemme, og de som av og til eller aldri snakker norsk i hjemmet. Resultatene viser at de som snakker norsk i hjemmet presterer vesentlig bedre enn de som ikke gjør det på lesetesten. (UDIR, 2001, s. 50)

2.2. Familiebakgrunn

Barn med ressurssterke foreldre i form av utdanning og inntekt, lykkes i gjennomsnitt bedre i skolesystemet enn barn fra mindre ressurssterke familier (Rauum, s.114, 2003). I Rauum sin forskningsrapport om oppvekstmiljø og utdanningsløp, pekes det på at foreldre med høyere utdanning enklere kan bistå barn med hjelp til skolearbeid. Det nevnes også at evnen til konsentrasjon, refleksjon, interesse og samarbeid er evner som går i arv, og samtidig som styrkes av sosial omgang (Rauum, s. 115, 2003). Dette tyder på at barn med ressurssterke foreldre påvirkes av å samhandle, prate og reflektere sammen med sine foreldre, noe som barn med mindre ressurssterke foreldre ikke vil ha lik mulighet til.

Samvariasjonen mellom foreldre og barn sin utdanningslengde presenteres i tall fra SSBs utdanningsregister fra 1993, hvor barn med mødre som har 13-16 års skolegang, går på skole i gjennomsnitt 3.7 år lengre på skole enn barn med mødre som har 7-9 års skolegang. Tilnærmet samme resultat har fedres lengde på utdanning på barnas utdanningsløp. (Rauum, s.119, 2003)

Ifølge boken "*Language in a Global World; Learning for better Cultural Understanding*" av OECD, finnes det en sterk sammenheng mellom leseferdigheter og innvandrerstatus. Rundt 40 % av innvandrerbarn i Tyskland oppnår ikke det mest basale nivået for leseferdigheter. For barn som ikke er innvandrere er det samme tallet 14 %. Innvandrere har generelt lavere utdanning og derfor kan det tyde på at innvandrerbarn også vil ha lavere sannsynlighet for å oppnå universitetsutdanning. (Mathä, Porpiglia, Sierminska, 2011, s. 14) & (OECD, 2012, s. 364-365)

2.3.Lærerkvalitet

I artikkelen "*Education Production Functions*", skriver Eric A. Hanushek blant annet at det er utført en del forskning for å se på effekten av ressursbruk rettet mot ulike skoler. Den generelle konklusjonen viser at det ikke er en tydelig sammenheng mellom hvor mye ressurser som brukes, men heller hvordan de brukes. Det mest signifikante resultatet er knyttet til lærerkvalitet. En studie gjennomført av Hanushek i 2011 peker på nettopp dette. En lærer blant de 25% dyktigste i et utvalg kunne generere så mye som $400 000 i økte inntekter aggregert over en klasse på 30 elever, sammenlignet med en gjennomsnittlig lærer. På motsatt side av skalaen finner Hanushek at en av de 10% svakeste lærerne kunne redusere aggregert inntekt med så mye som $800 000. Tidligere forskning indikerer altså at kvaliteten på læreren er det som ser ut til påvirke elevprestasjonene mest, men det skal nevnes at vi i denne oppgaven kun bruker leseferdigheter som et mål på elevprestasjoner. I vårt datasett er det også svært få variabler knyttet til lærerkvalitet, og de er meget svakt korrelert med leseferdigheter (se 8.1.8 i appendiks). Derfor har vi valgt å fokusere på andre faktorer. (Hanushek, 2020, s.167-169)

3. Økonometriske modeller og teori

3.1. Innledning

I dette kapittelet skal vi gjennomgå den økonometriske metoden vi skal anvende, nemlig minste kvadraters metode. I tillegg skal vi presentere ulike hypotesetester, og hvordan disse metodene kan anvendes på datamaterialet for å undersøke vår problemstilling.

3.2. Metode

Minste kvadraters metode (OLS) benyttes for å estimere de ukjente parameterne i en lineær regresjonsmodell. Metoden muliggjør det å estimere virkningen de uavhengige variablene X_i har på den avhengige utfallsvariabelen *Y (read).* Ved hjelp av OLS kan vi undersøke hvorvidt det finnes en sammenheng mellom prestasjoner i skolen og sosiodemografiske forhold. Vi tar utgangspunkt i en lineær regresjonsmodell. (Thomas, 2005, s. 266)

3.2.1. Formel: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \varepsilon$

Med denne modellen kan vi se hvordan endringer i de uavhengige variablene X_i påvirker *Y*. Her er det viktig å understreke at den faktiske verdien på *Y* ikke nødvendigvis er lik forventet verdi på *Y*. Videre forutsetter vi lineær årsakssammenheng mellom den avhengige utfallsvariabelen *Y* og en eller flere uavhengige variabler X_i. (Thomas, 2005, s.356)

Restleddet ε er gitt som differansen mellom forventningsverdien *E(Y)* og faktisk verdi på *Y*. Leddets verdi avhenger av de uforklarte og ikke-målbare faktorene som spiller inn på den forventede verdien til populasjonen. Jo større restleddet er, jo større er de uforklarte og ikke-målbare faktorene (Thomas, 2005, s. 356). Den klassiske regresjonsmodellen vi tar i bruk krever sterke forutsetninger knyttet til de uavhengige variablene X_i og restleddet ε før vi bruker den i vår analyse av datamaterialet:

Forutsetninger til de uavhengige variablene (Thomas, 2005, s.357-359):

- De uavhengige variablene kan ikke være stokastiske i motsetning til avhengig variabel *(Y)*, ettersom verdien på *X* ikke er tilfeldig eller blir determinert av andre variabler.
- Om *X*-verdiene holder seg konstante over flere målinger, medgir ikke dette at *Y*-verdien gjør det. Dette fordi variasjonen i restleddet ε ikke kontrolleres av størrelsen på *X*-verdiene.
- Hvis antall undersøkte *n* øker, vil ikke nødvendigvis variansen av *X*-verdiene øke betraktelig. I oppgaven har vi ikke tidsseriedata, noe som kunne gjort denne siste forutsetningen ustabil.

Forutsetninger til restleddet (Thomas, 2005, s. 359-361):

- $E(\varepsilon_i) = 0$: Forventet verdi til restleddet settes lik null, som i korte trekk betyr at avviket mellom estimerte verdier og observerte verdier er lik null.
- $V(\varepsilon_i) = \sigma^2 < \infty$: Residualene har lik varians for alle *X*-er, noe som betyr at restleddet er homoskedastisk.
- $Cov(\varepsilon_i, \varepsilon_j) = 0$: Korrelasjonen mellom to ulike restledd er null. De er altså uavhengige av hverandre.
- $\varepsilon_i \sim N(0, \sigma^2)$: Restleddet er normalfordelt, med et gjennomsnitt på 0.

Videre skal vi finne den estimerte regresjonslinjen for utvalget vårt ved hjelp av OLS. Da minimerer vi summen av variansen, som er det kvadrerte avviket mellom estimert verdi på *Y (read)* og observert verdi $\hat{Y}$ (Thomas, 2005, s.274-276). Ved minimering får vi parameterne *a* og *b* i følgende predikert regresjonslinje:

3.2.2. Formel: $\hat{Y} = a + bX_i$

Vi bruker determinasjonskoeffisienten R^2 for å undersøke hvor godt regresjonslinjen beskriver datasettet. Koeffisienten angir hvor stor andel av total endring i *read* (*Y*) som kan forklares av de uavhengige variablene (X_i) (Thomas, 2005, s. 275-277). Dette medfører at forklaringskraften R^2 øker, dersom flere variabler inkluderes i modellen. For å unngå misvisende resultater som følger av dette, er en mulig løsning å benytte den justerte R^2. Dette lar seg derimot ikke korrigere for når vi velger å benytte robuste standardavvik, slik beskrevet i 8.2 (Thomas, 2005, s.277).

3.2.3. Formel: $R^2 = \frac{SSE}{SST}\left(= \frac{b^2 \sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2}\right), \qquad 0 < R^2 < 1$

3.3. Hypotesetesting

En hypotesetest er en statistisk testmetode som anvendes for å kunne si noe om egenskapene til en populasjon. For å teste en hypotese formulerer man en nullhypotese *(H_0)* og en alternativhypotese *(H_A).* Dersom man har grunnlag til å forkaste nullhypotesen gir det støtte til alternativhypotesen. (Thomas, 2005, s.258)

I denne oppgaven benyttes hypotesetester for å kunne avgjøre om regresjonsestimatorene er forskjellige fra null. Dersom estimatoren blir signifikant forskjellig fra null, gir det grunnlag for å kunne forkaste nullhypotesen. Signifikansnivået er et mål på hvor stor usikkerhet en er villig til å akseptere i modellen. Det uttrykker sannsynligheten for å forkaste en gyldig nullhypotese, og vi legger til grunn et signifikansnivå på 5%. (Thomas, 2005, s. 370)

For å kontrollere om koeffisientene i regresjonen har signifikant effekt på den avhengige variabelen, benytter vi oss av t-testen. Denne testen undersøker om den gjennomsnittlige verdien i et normalfordelt datasett er signifikant forskjellig fra nullhypotesen. I våre tester benytter vi signifikansnivå på 5% som for en ensidig test av store utvalg har kritisk verdi $t_{0.05} = \pm 1.645$. Når nullhypotesen er at en variabel ikke har en signifikant påvirkning blir testobservatoren gitt som $TS = \frac{b_j}{s_{b_j}}$, der b_j er estimatet for den ukjente β og s_{b_j} er standardavviket til estimatet. Dersom absoluttverdien til testobservatoren overstiger kritisk verdi, forkaster vi nullhypotesen. (Thomas, 2005, s. 397)

En F-test kan benyttes for å teste om multiple hypoteser har en signifikant effekt på den avhengige variabelen. Dette kan være relevant for å avgjøre om man skal inkludere variabler i modellen. F-testen baserer seg på endringer i summen av kvadrerte residualer (SSR) mellom en restriktiv og en urestriktiv modell. Testobservatoren er F-fordelt med *(h, n – k)* frihetsgrader, der *h* er antall restriksjoner, *n* er utvalgsstørrelse og *k* er antall parametere i likningen. (Thomas, 2005, s. 419)

3.3.1. Formel: $TS = \frac{(R_U^2 - R_R^2)/h}{(1 - R_U^2)/(n-k)} \sim F_{h,n-k}$

3.4. Korrelasjon

Korrelasjon er et statistisk mål på sammenhengen mellom to eller flere variabler. Vi skiller mellom positiv og negativ korrelasjon; positiv dersom x øker når y øker, og negativ dersom x reduseres når y øker. For å kunne sammenligne korrelasjon bruker vi korrelasjonskoeffisienten (ρ), som ikke påvirkes av måleenhetene og tar verdi mellom -1 og 1. Dersom koeffisienten er lik 1, har vi perfekt positiv korrelasjon. Hvis den er -1, har vi perfekt negativ korrelasjon. Er korrelasjonskoeffisienten tilnærmet lik 0, betyr det at størrelsene er uavhengige, altså at det ikke er noen sammenheng mellom variablene. (Thomas, 2005, s. 194-195)

3.4.1. Formel: $R = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}, \qquad -1 \leq R \leq 1$

Det er også viktig å presisere at korrelasjon ikke nødvendigvis indikerer kausalitet, altså at X forårsaker Y. En variabel Z kan være årsaken til at det er en korrelasjon mellom X og Y. Når vi har en korrelasjon som ikke er kausal, kalles det for en spuriøs sammenheng (Thomas, 2005, s. 258). Et problem vi kan støte på i våre regresjoner er multikollinearitet, som betegner graden av lineær sammenheng mellom forklaringsvariablene. En konsekvens av multikollinearitet er at standardavviket til koeffisientene i regresjonen kan øke, slik at estimatene blir svært sensitive på små endringer. Vi har derfor undersøkt korrelasjonen mellom aktuelle forklaringsvariabler, slik vist i korrelasjonsmatrisen (Tabell 8.1.7). (Thomas, 2005, s. 402)

4. Datamateriale

4.1. Introduksjon av datamateriale

I dette kapittelet skal vi presentere datamaterialet fra den internasjonale leserundersøkelsen *Progress in International Reading Literacy Study*. PIRLS blir gjennomført hvert femte år og undersøker leseferdigheter for elever på fjerde og femte trinn. Elevene svarer på spørsmål om deres lesevaner, hvordan de opplever undervisning og hvordan deres motivasjon til lesing er. Undersøkelsen tar også for seg spørsmål knyttet til foreldres og skolelederes tilrettelegging for elevenes lesing. Vi skal med bakgrunn i denne dataen utforske hva det er som påvirker elevenes lesekompetanse. (Utdanningsdirektoratet, 2017)

Leseheftene som brukes i PIRLS er fagtekster og litterære tekster med spørreskjemaer hvor deltakeren selv må gjengi hva som er lest. PIRLS legger vekt på forståelsen for hvorfor man leser, at man skal huske det man har lest, og hvilke holdninger man har til lesing (lesesenteret, Wagner, 2018). I denne oppgaven har vi brukt datamateriale fra Norge og Tyskland, basert på individuelle rapportert levert av hvert av landene.

4.2. Datasettets innhold

I bacheloroppgaven ønsker vi å undersøke hvilke forhold som påvirker elevers leseferdigheter, representert ved vår avhengige variabel *read*. Dette er en kontinuerlig variabel som måler elevenes poengsum på lesetesten. Resultatene er standardiserte, noe som gjør det mulig å sammenligne på tvers av land. Variablene er ytterligere beskrevet i tabell 8.1.4 i appendiks.

4.3. Husholdningskarakteristika

To sentrale kontrollvariabler er *income* og *par_edu* som henholdsvis karakteriserer ulike nivåer på foreldres årlige inntekt målt i amerikanske dollar, og høyeste oppnådde utdanning blant begge foreldre. Variabelen *books_home* angir et kategorinivå for antall bøker i hjemmet. Disse variablene er indikatorer for læringsmiljøet i hjemmet og sosioøkonomisk status for den enkelte elev. (timssandpirls, s. 38, 2001)

4.4. Elevkarakteristika

Når vi videre omtaler innvandrere i oppgaven henviser vi til elevene som kategoriseres som *par_not_born*, altså elever med foreldrene som ikke er født i det respektive landet. Alternativet vil være å bruke variabelen *not_born*, som betegner elever som selv ikke er født i det respektive landet. Fordelen med *par_not_born* er at den i stor grad vil inkludere både første- og andregenerasjons innvandrere. Variabelen *early_ability* betegner leseferdigheter på et tidligere stadium i skoleløpet. Dette er en interessant kontrollvariabel for å undersøke om tidlig utvikling av leseferdigheter påvirker poengsummen på leserundersøkelsen. Kategorivariabelen *pct_abroad* defineres som prosentandel av elevene på skolen som er født i utlandet, og inkluderes i modellen for å undersøke om dette påvirker læringsmiljøet. Dummyvariablene *girl* og *grmny* beskriver henholdsvis kjønn og om eleven bor i Norge eller Tyskland.

4.5. Deskriptive data

4.5.1. Tabell – Deskriptiv data for avhengig variabel, read

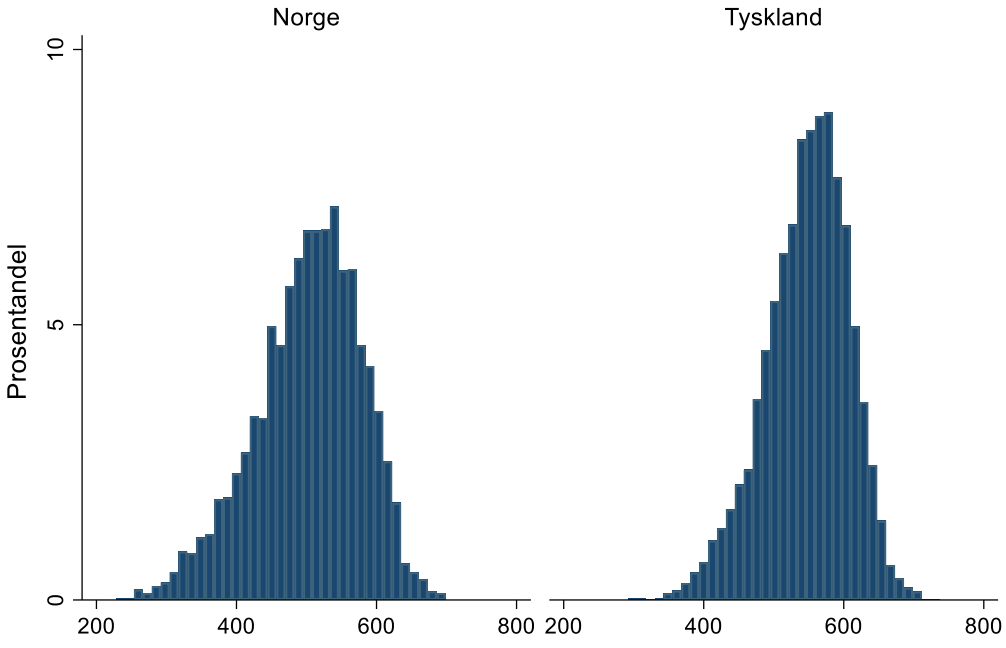
	Alle		Innvandrere		Ikke-innvandrere	
	Norge	Tyskland	Norge	Tyskland	Norge	Tyskland
Gjennomsnitt	499.23	538,53	463.79	495.92	502.77	550.18
Standardavvik	78.12	63,79	83.41	63.05	76.69	58.84
Minimum	228.07	289.44	234.89	289.44	228.06	296.48
Maksimum	695.87	724.22	673.03	689.33	695.87	724.22
Antall observasjoner	3355	7055	305	515	3050	5540

Verdier hvor det mangler data for innvandrerstatus er utelatt

Vi skal se på forskjeller i leseferdigheter mellom innvandrere og ikke-innvandrere i Norge og Tyskland ved hjelp av deskriptiv statistikk. Fra de deskriptive dataene for vår avhengige variabel *read*, kan vi se at leseferdighetene totalt sett er bedre i Tyskland enn i Norge. Tyskland gjør det generelt sett 7,3% bedre enn Norge i leserundersøkelsen.

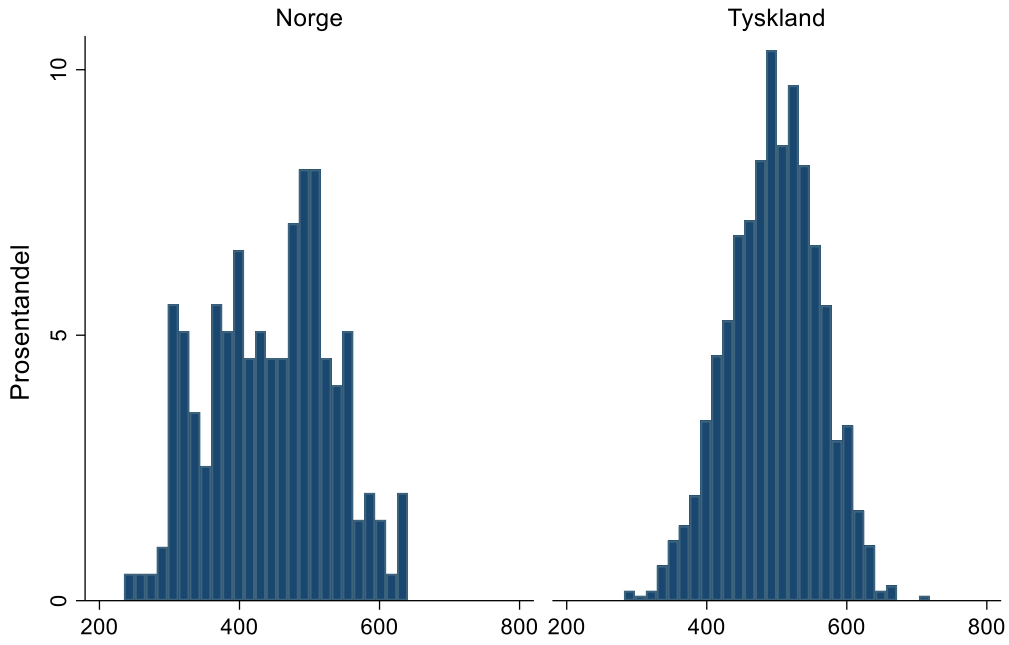
Vi skal nå rette oss inn mot det som er kjernen av problemet vårt; om det eksisterer et signifikant prestasjonsgap mellom innvandrerelever og øvrige i de to landene. Innvandrere i Norge presterer i gjennomsnitt 7,7% svakere enn øvrige elever. Dette er tall som noenlunde sammenfaller med UDIR-tall fra 2001, hvor øvrige elever presterte 10,7% bedre enn innvandrerelever. Innvandrere i Tyskland presterer 9,8% svakere enn øvrige tyske elever. Den deskriptive analysen viser altså at innvandrere i Tyskland gjør det svakere i forhold til øvrige elever, enn innvandrere i Norge gjør det i forhold til øvrige. Denne forskjellen er ganske liten og vi vil i kapittel 5 vise, at den ikke er signifikant. Vi ser av statistikken at det eksisterer et prestasjonsgap mellom innvandrer - og ikke-innvandrerelever. Hva som forklarer årsakene i denne dataen, skal vi se nærmere på i kapittel 5. Tabellen viser at det generelt er relativt store forskjeller i de to landene mellom de svakeste og beste elevene. Vi kan observere fra standardavviket at spredningen generelt er større i Norge. Det er interessant å se at de beste tyske elevene (724,22) presterer vesentlig høyere enn de beste norske (695,87). I tillegg er bunnivået blant norske elever (228,07) vesentlig lavere enn for tyske elever (289,44).

4.5.2. Tabell – Leseferdigheter for ikke-innvandrere



Variabel: *read* if *par_not_born = 0*

4.5.3. Tabell – Leseferdigheter for innvandrere



Variabel: *read* if *par_not_born = 0*

Histogrammene illustrerer ulikhetene i leseferdigheter blant norske og tyske innvandrere, samt ikke-innvandrere fra begge land. Ut ifra histogrammene kan man observere at nordmenn ser ut til å prestere på et lavere nivå med en større spredning i resultatene enn tyskerne. Ulikhetene i spredning kan ha en sammenheng med at det er vesentlig flere respondenter i Tyskland.

4.5.4. Tabell – Deskriptiv statistikk for uavhengige variabler

	Norge		Tyskland	
	Gjennomsnitt	Standardavvik	Gjennomsnitt	Standardavvik
not_born	0.0909	-	0.2147	-
par_not_born	0.0584	-	0.1538	-
girl	0.4810	-	0.4931	-
income	4.0641	1.5501	3.1503	1.6112
books_home	4.0329	1.04880	3.4683	1.1997
par_edu	1.9509	1.0555	2.2613	1.0569
early_ability	2.6194	0.9467	2.2644	0.9557
pct_abroad	1.1401	0.3927	1.5075	0.7962
speak_testlang_home	1.1118	0.3573	1.1519	0.4053

Vi har valgt å utelate standardavviket for dummyvariablene da de kun kan ta inn to verdier.

Gjennomsnittlig inntektsnivå i Norge er 4,0641 som er innenfor inntektsintervallet $40.000-49.999, og 3,1503 i Tyskland som er innenfor inntektsintervallet $30.000-39.999. Bøker i hjemmet (books_home) er sett på som en indikator på hjemmets tilrettelegging for at barn skal bli motivert til å lese. I et norsk hjem er det gjennomsnittlige antall bøker 101-200, mens det i Tyskland er 26-100. (timssandpirls.bc.edu, s. 38, 2001)

4.5.5. Tabell – Gjennomsnittlig poengsum på leserundersøkelsen for ulike antall bøker i hjemmet

	Norge		*Tyskland*	
	Gjennomsnitt	Standardavvik	Gjennomsnitt	Standardavvik
Read_books_home1	444.5338	81.0312	487.4992	61.086
Read_books_home2	455.4217	78.9697	504.9214	63.3142
Read_books_home3	477.9229	75.8647	530.7907	58.6502
Read_books_home4	499.9065	77.5634	549.1556	58.0066
Read_books_home5	523.8672	70.9402	570.6928	56.2328

Statistikken viser at for hvert intervall av bøker i hjemmet presterer tyske elver bedre enn norske. Det er betydelig større forskjell mellom gjennomsnittlige leseferdigheter *(read)* blant norske og tyske elever i intervall 3 (26-100), enn ved færre bøker i hjemmet (0-10) presentert i intervall 1. Tyske elever gjør det generelt bedre enn norske elever til et hvert antall bøker i hjemmet.

4.5.6. Tabell – Gjennomsnittlig poengsum på leserundersøkelsen for elever i husholdninger med ulik inntekt

	Norge		*Tyskland*	
	Gjennomsnitt	Standardavvik	Gjennomsnitt	Standardavvik
Read_Income1	468.1251	87.1013	506.7701	64.2025
Read_Income2	487.1942	82.0739	524.6119	60.5281
Read_Income3	487.6753	79.8624	541.9136	60.4703
Read_Income4	493.4376	75.676	554.4721	56.8186
Read_Income5	509.0219	71.6061	570.7674	54.6545
Read_Income6	530.0115	68.8477	571.8558	53.9884

Tyske elever fra den fattigste inntektsgruppen gjør det gjennomsnittlige bedre enn hva norske elever i gjennomsnitt gjør det frem til og med inntektsnivå 5. Vi ser også at for hvert gitte norske inntektsnivå, gjør tyske elever det i gjennomsnitt bedre enn norske elever på hvert tilsvarende inntektsnivå. Norske elever som tilhører inntektsgruppe 2 ($20,000-29,999) gjør det tilnærmet likt som norske elever i inntektsgruppe 3 ($30,000-39,999). Samme fenomen ser vi også i Tyskland blant elever som tilhører husholdninger som inngår i inntektsgruppe 5 ($50,000-59,999) og 6 ($60,000 eller mer). Spredningen er også mye større i Norge enn i Tyskland noe som tilsier at norske elever er mer ujevne i prestasjonene enn tyske.

4.6. Styrker og svakheter til datasettet

Regresjonen vår består av flere kategorivariabler, hvor det er ulike intervaller mellom nivåene. Variabelen *income* som sier hvilken inntektsgruppe husholdningen tilhører, har for eksempel ulike intervaller mellom inntektsgruppene. Dette må man ta hensyn til under tolkningen av regresjonen, der en mulighet er å dele kategorivariabler inn i dummyvariabler. Vi velger heller å beholde kategorivariablene som de er, slik at tolkningen av koeffisientene i regresjonen betraktes som hopp mellom kategorinivåer.

Det er viktig å redegjøre for at på tidspunktet undersøkelsen ble gjort i 2001 var 6,3% av Norges befolkning innvandrere, hvor dansker, svensker og pakistanere utgjorde mesteparten av disse (ssb.no, 2001). Det vil være rimelig å anta at dansker og svensker vil ha en fordel på testen, da deres språk er relativt likt norsk, som er språket undersøkelsen blir gjennomført på. I Tyskland er de fleste innvandrere tyrkere (worldatlas.com, 2019). Dette er mennesker som har et morsmål svært ulikt tysk. Dette kan tyde på at innvandrere i Norge har en kortere vei til å lære seg språket testen gis ut på, enn innvandrere i Tyskland har.

Det kunne også vært interessant å undersøke noen variabler som ikke inngår i dette datasettet. Det kunne for eksempel vært å ha en variabel for om eleven går på privat eller offentlig skole, og om dette har noen effekt på leseferdighetene. Gode variabler på lærerkvalitet ville vært interessant å se på i henhold til Hanushek sine resultater nevnt i kapittel 2. I tillegg bruker vi kun resultatene fra lesetester, og det kunne vært interessant å se på hvordan disse elevene gjør det i andre fag om matematikk og naturfag.

5. Regresjonsanalyse

5.1. Innledning

I den deskriptive analysen observerte vi at de tyske elevene i gjennomsnitt gjør det bedre enn de norske elevene på lesetestene. Vi så også at innvandrerelever generelt sett gjør det svakere enn øvrige elever både i Norge og i Tyskland. Fra kapittel 4 har vi i tillegg sett at prestasjonsgapet mellom innvandrere og øvrige elever er noe større i Tyskland enn i Norge. I regresjonsanalysen ønsker vi å undersøke hvilke faktorer som påvirker leseferdighetene, da særlig innvandrerstatus. Videre ønsker vi å undersøke om det eksisterer et prestasjonsgap mellom innvandrere og ikke-innvandrere, og sammenlikne dette i Norge og Tyskland. Det blir da relevant å se på hvilken effekt kjønn og andre sosiodemografiske variabler har å si på leseferdighetene.

Regresjonsanalysen gjennomføres ved hjelp av minste kvadraters metode i *Stata*. Vi velger en lin-lin-funksjonsform i regresjonsmodellen, da dette er det mest vanlige innenfor feltet. Den avhengige variabelen er leseferdigheter (*read)*. Siden leseferdighetene er målt i poeng og standardiserte mellom land er det fornuftig å bruke de faktiske verdiene. De estimerte koeffisientene i regresjonen forteller hvor mye leseferdigheter endres dersom den respektive variabelen endres med en enhet og alle de andre variablene er uendret. Funksjonsformen vi tar utgangspunkt i tar form som følgende

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \varepsilon$$

I regresjonen har vi valgt å slå sammen datasettene for Norge og Tyskland, og generert dummyvariabelen *grmny* som tar verdien *0* dersom eleven bor i Norge, og verdien *1* dersom eleven bor i Tyskland. Vi estimerer de fem følgende modellene.

$$(1) \quad read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \varepsilon$$

$$(2) \quad read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \rho_1 pnb\_girl_i + \varepsilon$$

$$(3) \quad read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \delta_3 grmny_i + \varepsilon$$

$$(4) \quad read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \delta_3 grmny_i + \rho_2 girl\_grmny_i + \delta_3 pnb\_grmny_i + \varepsilon$$

$$(5) \quad read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \delta_3 grmny_i + \rho_2 girl\_grmny_i + \beta_2 early\_ability_i + \beta_3 books\_home_i + \beta_4 par\_edu_i + \beta_5 pct\_abroad_i + \varepsilon$$

5.2.Empiriske resultater

	m1	*m2*	*m3*	*m4*	*m5*
VARIABLES	**read**	**read**	**read**	**read**	**read**
par_not_born	-29.991	-30.373	-37.519	-37.057	-29.631
	(2.583)	*(3.554)*	*(2.479)*	*(7.676)*	*(2.811)*
girl	15.361	15.281	15.017	19.483	14.532
	(1.522)	*(1.606)*	*(1.415)*	*(2.758)*	*(2.602)*
income	7.245	7.244	11.059	11.051	4.946
	(0.463)	*(0.463)*	*(0.454)*	*(0.454)*	*(0.555)*
pnb_girl		0.762			
		(5.044)			
grmny			54.450	57.931	70.683
			(1.661)	*(2.364)*	*(2.371)*
girl_grmny				-6.985	-6.589
				(3.176)	*(3.169)*
pnb_grmny				-0.544	
				(8.054)	
early_ability					15.890
					(0.822)
books_home					12.475
					(0.783)
par_edu					-10.915
					(0.856)
pct_abroad					-4.058
					(1.211)
Constant	499.7	499.8	452.5	450.3	411.0
	(2.034)	*(2.060)*	*(2.543)*	*(2.815)*	*(5.496)*
Observations	8,051	8,051	8,051	8,051	6,117
R-squared	*0.067*	*0.067*	*0.193*	*0.194*	*0.312*

Standardavvik i parenteser

5.3. Modell 1 – Innvandrerstatus og kjønn påvirker leseferdighetene

$$(1) \qquad read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \varepsilon$$

Modell 1 er en svært restriktiv modell som ser på den generelle effekten av innvandrerstatus og kjønn på leseprestasjonene. Det er en enkel lineær regresjon som inneholder interessevariablene *par_not_born* og *girl*, i tillegg til kontrollvariabelen *income* som korrigerer for husholdningens inntekt. Resultatene vil være estimater for et kombinert utvalg av både tyskere og nordmenn, da vi foreløpig ikke har tatt høyde for landtilhørighet i modellen. Alle observasjoner hvor det mangler data er utelatt, noe som gjør at antall observasjoner er merkbart lavere enn antallet av elever som deltok i undersøkelsen.

Vi kan illustrere hvordan de estimerte koeffisientene påvirker konstantleddet, ved å sette verdiene fra regresjonen inn i modellen.

$$read_i = 499.7 - 29.99\, par\_not\_born_i + 15.36 girl_i + 7.25 income_i$$

Variabelen *par_not_born* har en negativ koeffisient på -29.99. Dette tolkes slik at det å være innvandrer vil påvirke leseferdighetene negativt med omtrent 30 poeng, gitt at de andre variablene holdes konstante. Tilsvarende kan vi se fra regresjonen at jenter får omtrent 15 poeng mer enn gutter. Inntekt har også en del å si, hvor vi ser at en økning i inntektsgruppe i gjennomsnitt vil medføre høyere leseferdigheter på 7 poeng. Dette er en kategorivariabel for 6 inntektsnivåer, som da kan ta verdi fra 0 til 5. De estimerte koeffisientene viser hvordan variablene påvirker konstantleddet på 499.7. På denne måten kan vi for eksempel estimere leseferdigheter for en innvandrerjente, hvor familien er i inntektsgruppe 1, til å bli 485 poeng.

$$\left(read = 499.7 - 29.99 \cdot 1 + 15.36 \cdot 1 + 7.25 \cdot 0 = 485.07\right)$$

Regresjonsmodellen har en relativt lav forklaringskraft på 6.7 %, men poenget her er heller å se på en generell effekt av interessevariablene på leseferdigheter enn å oppnå høy forklaringskraft. Innvandrerstatus, kjønn og inntektsgruppe forklarer altså omtrent 6.7% av variasjonen i leseferdighetene.

Vi ønsker å teste om effekten av innvandrerstatus på leseferdighetene er signifikant forskjellig fra null ved hjelp av en t-test. Nullhypotesen er at innvandrerstatus ikke påvirker leseferdighetene, mens alternativhypotesen er at innvandrerstatus påvirker leseprestasjonene i negativ grad.

$$H_0 : \delta_1 = 0, \qquad H_A : \delta_1 < 0$$

Vi velger en ensidig t-test med signifikansnivå på 5% som for store utvalg har kritisk verdi $t_{0.05} = \pm 1.645$. Testobservatoren er gitt ved: $TS = \frac{b_j}{s_{b_j}} = \frac{-29.991}{2.553} = -11.75$

Absoluttverdien til testobservatoren overstiger kritisk verdi og vi må forkaste nullhypotesen. Vi kan konkludere med at innvandrerstatus påvirker leseferdighetene negativt innenfor et 5% signifikansnivå.

Vi ønsker også å teste for kjønnseffekten på leseferdigheter, med nullhypotesen som sier at det ikke er noen forskjell i leseferdigheter mellom kjønn. Alternativhypotesen er at jenters leseferdigheter er på et høyere nivå enn gutters.

$$H_0: \quad \delta_2 = 0, \qquad H_A: \quad \delta_2 > 0$$

$$TS = \frac{15.361}{1.522} = 10.09$$

Testobservatoren overstiger TS på et 5% signifikansnivå. Vi kan dermed forkaste nullhypotesen og konkludere med at jenter presterer bedre enn gutter.

5.4. Modell 2 – Kjønnseffekten for innvandrere og ikke-innvandrere

$$(2) \qquad read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \rho_1 pnb\_girl_i + \varepsilon$$

Modell 2 bygger videre på modell 1, men det tilføyes et interaksjonsledd mellom dummyvariablene *par_not_born* og *girl*. Det tillates altså at prestasjonsforskjeller mellom kjønn også kan avhenge av innvandrerstatus. Dette gjøres for å se på hvorvidt kjønnseffekten er ulik blant henholdsvis innvandrere og ikke-innvandrere. Referansekategorien er gutt som er ikke-innvandrer.

Ut ifra resultatene fra regresjonen observerer vi at *pnb_girl* har en koeffisient på 0.762, noe som isolert sett betyr at kjønnseffekten er noe større for innvandrere. Denne effekten er liten, og vi ønsker å sjekke om den er statistisk signifikant. Vi tester om prestasjonsforskjellene mellom kjønn for innvandrere og ikke-innvandrere er signifikant ulike. For å finne ut av dette kan vi utføre en t-test på interaksjonsvariabelen *pnb_girl*, der nullhypotesen er at kjønnseffekten på prestasjoner er uavhengig av innvandrerstatus. Alternativhypotesen er at kjønnseffekten på prestasjoner er større for innvandrere enn ikke-innvandrere. I *Stata* observerer vi at t-verdien er lik 0.15, noe som ikke er signifikant forskjellig fra null på et 5% signifikansnivå. Vi kan ikke forkaste nullhypotesen og har ikke grunnlag for å si at kjønnseffekten er avhengig av innvandrerstatus. Siden eneste forskjellen fra

modell 1 er inkludering av en ikke-signifikant interaksjonsvariabel ser vi at heller ikke forklaringskraften endres fra 6.7%.

5.5. Modell 3 – Prestasjonsgap i Norge og Tyskland

$$(3) \qquad read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \delta_3 grmny_i + \varepsilon$$

Modell 3 ser på effekten av hvilket land (Norge eller Tyskland) eleven kommer fra har på leseferdighetene. Modellen legger opp til å undersøke om det finnes et prestasjonsgap mellom Norge og Tyskland. Denne forskjellen uttrykkes med interaksjonsleddet *grmny*, som tillater at leseferdighetene kan være forskjellig mellom norske og tyske elever. Koeffisienten for dette leddet har en verdi på 54.45. Ifølge estimatene fra denne modellen vil altså en tysk elev prestere rundt 54 poeng bedre enn en norsk elev når man holder de andre variablene konstant. Når vi nå har korrigert for om eleven er norsk eller tysk, har det medført at modellens forklaringskraft har blitt betraktelig høyere.

Vi ønsker å teste om dette prestasjonsgapet mellom Norge og Tyskland er signifikant ved hjelp av en ensidig t-test. Nullhypotesen sier at leseferdighetene er like i landene, mens alternativhypotesen sier at leseferdighetene er høyere i Tyskland enn i Norge.

$$H_0: \quad \delta_3 = 0, \qquad H_A: \quad \delta_3 > 0, \qquad TS = \frac{54.45}{1.661} = 32.78$$

T-verdien overstiger i dette tilfellet kritisk verdi for et 5% signifikansnivå og vi må dermed forkaste nullhypotesen. Vi konkluderer med at leseferdighetene er høyere i Tyskland enn i Norge.

Dersom vi sammenligner modell 3 med de to foregående, ser vi at koeffisienten til *par_not_born* har blitt ytterligere redusert (fra -30 til -38), mens koeffisienten til *girl* er tilnærmet uendret. Innvandrereffekten ser altså ut til å være enda større når man tar hensyn til at det kan være forskjeller mellom landene, mens kjønnseffekten virker å være temmelig robust på tvers av landene. Dette ønsker vi derfor å undersøke ytterligere i modell 4.
Modellen kan omformuleres til en modell for Norge og en for Tyskland. Dersom eleven er tysk, tar dummyvariabelen *grmny* verdien 1 og koeffisienten δ_3 inkluderes som et konstantledd. Dersom eleven derimot er norsk, tar *grmny* verdien 0, og koeffisienten inkluderes ikke.

Tyskland: $read_i = (\alpha + \delta_3) + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \varepsilon$

Norge: $read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \varepsilon$

5.6. Modell 4 – Innvandrer- og kjønnseffekt i Norge og Tyskland

$$(4) \quad read = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \delta_3 grmny_i + \rho_2 girl\_grmny_i + \delta_3 pnb\_grmny_i + \varepsilon$$

Modell 4 tillater at innvandrereffekten og kjønnseffekten på leseferdighetene avhenger av om eleven er norsk eller tysk. På bakgrunn av observasjonene i modell 3 ønsker vi å undersøke om disse effektene er ulike mellom landene. Vi har generert interaksjonsvariablene *girl_grmny* og *pnb_grmny*, der referansegruppen er gutt og ikke-innvandrer i Norge.

Girl_grmny er et interaksjonsledd mellom kjønn og land. Dette interaksjonsleddet skal analysere hvorvidt det er en statistisk signifikant forskjell mellom hvor gode leseferdigheter gutter og jenter har i Norge og Tyskland. Ut ifra estimatene fra regresjonen ser vi at en norsk jente i snitt gjør det nesten 20 poeng bedre enn en norsk gutt. Vi ser fra koeffisienten for *girl_grmny* at forskjellen mellom gutter og jenter i Tyskland er rundt 7 poeng mindre enn i Norge. Det vil si at en tysk jente i snitt gjør det ca. 13 poeng bedre enn en tysk gutt. Vi ønsker å kontrollere for om denne effekten er signifikant forskjellig fra null med en t-test. Nullhypotesen er at forskjellen i leseferdigheter mellom kjønn er lik i Norge og Tyskland. Fra *Stata* observerer vi at t-verdien er lik -2.20, hvis absoluttverdi overstiger kritisk verdi for et 5% signifikansnivå. Vi må altså forkaste nullhypotesen og kan konkludere med at forskjellen mellom kjønn er mindre i Tyskland enn i Norge.

Interaksjonsleddet *pnb_grmny* skal analysere om det finnes en forskjell i leseferdigheter mellom norske og tyske innvandrere og ikke innvandrere. Med en koeffisient på beskjedne -0.5 tilsier det at innvandrereffekten er noe mindre i Tyskland enn i Norge. Vi ser fra t-verdien -0.07 i *Stata* at effekten ikke er signifikant forskjellig fra null og har altså ikke grunnlag for å si at innvandrereffekten er ulik mellom landene. Det vil altså si at forskjellen i leseferdigheter mellom en tysk innvandrer og en tysk ikke-innvandrer er omtrent like stor som forskjellen mellom en norsk innvandrer og ikke-innvandrer.

Sammenliknet med modell 3 kan vi observere at standardavviket til koeffisientene til *par_not_born* og *girl* øker betraktelig. Det skapes dermed større usikkerhet om hvor presist estimatene er. Årsaken til dette er trolig korrelasjon mellom *pnb_girl* og *pnb_germny*, noe vi kan se i korrelasjonsmatrisen 8.1.7 i appendiks. Regresjonen har en forklaringskraft med R-squared lik 19.4%, noe som er svært likt modell 3. Dette skyldes at det kun er tilført interaksjonsvariabler mellom variabler som allerede er inkludert.

5.7. Modell 5 – Sosiodemografiske variabler påvirker leseferdighetene

$$(5) \quad read_i = \alpha + \delta_1 par\_not\_born_i + \delta_2 girl_i + \beta_1 income_i + \delta_3 grmny_i + \rho_2 girl\_grmny_i + \beta_2 early\_ability_i + \beta_3 books\_home_i + \beta_4 par\_edu_i + \beta_5 pct\_abroad_i + \varepsilon$$

I modell 5 ønsker vi å undersøke om en rekke sosiodemografiske variabler påvirker leseferdighetene. Vi har tilført en rekke kontrollvariabler som korrigerer for tidlige leseevner, antall bøker i hjemmet, foreldrenes utdanning og innvandrerandel i klassen. Disse er alle kategorivariabler og tolkningen av en koeffisient for en slik variabel vil være at dersom du går opp et nivå i kategorien svarer koeffisienten til endringen i estimerte leseferdigheter. Denne endringen er gjennomsnittlig endring mellom alle nivåene i kategorivariabelen. Her har vi altså en modell med få restriksjoner. Da variabelen *pnb_grmny* ikke viste seg å ha en signifikant påvirkning i modell 4, har vi videre valgt å utelate den. I kapittel 2 så vi at det er en signifikant sammenheng mellom hvilket språk det snakkes i hjemmet og hva slags poengsum elevene oppnår. Variabelen *speak_testlang_home* var derfor av interesse som kontrollvariabel, men *vi* har valgt å utelate den fordi den er sterkt korrelert med *par_not_born.* Dette kan potensielt gi et problem med multikollinearitet.

Modell 5 viser effekten en rekke sosiodemografiske variabler har på leseferdighetene. Det er denne modellen som gir høyest forklaringskraft med R-squared på 31.2 %. Den tar inn både interaksjonsledd og kontroll – og interessevariabler. Denne modellen viser oss, at det er en sterk sammenheng mellom leseferdigheter, hvor foreldrene er født, kjønn, og om man bor i Norge eller Tyskland. Den viser dessuten at tidlige leseferdigheter, antallet av bøker i hjemmet og husstandsinntekt har stor betydning. Vi ser og at en økning av antall innvandrere på skolen gir svakere leseferdigheter blant elevene.

Vi ønsker å teste om kontrollvariablene vi har tilført i modell 5 samlet sett har en signifikant effekt på leseferdighetene. Dette gjør vi ved hjelp av en F-test, der vi har en restriktiv modell uten kontrollvariablene og en urestriktiv modell med kontrollvariablene.
Nullhypotesen er $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, med alternativhypotesen $H_A : \beta_2 = \beta_3 = \beta_4 = \beta_5 \neq 0$.
Modellenes verdi for forklaringskraft henter vi fra *Stata* og vi får følgende testobservator:

$$TS = \frac{(0.3116 - 0.1938)/4}{(1 - 0.3116)/(6117 - 10)} = 261.26$$

Kritisk verdi for et 5% signifikansnivå i en F-fordeling: $TS_{4,6117-10} = 2.3719 < 261.26$

Vi ser at testobservatoren overstiger kritisk verdi med god margin, og forkaster dermed nullhypotesen om at kontrollvariablene ikke har noen effekt på leseferdighetene.

Variabelnavn	*t-verdi*
early_ability	19.32
books_home	15.93
par_edu	-12.75
pct_abroad	-3.35

Tilsvarende kan vi ut ifra t-verdiene til kontrollvariablene observere at de også enkeltvis har en signifikant effekt. Ut ifra tabellen ser vi at alle t-verdiene overstiger kritisk verdi for et 5% signifikansnivå.

Estimatene viser at dersom du går fra en inntektsgruppe til en høyere, vil dette påvirke leseferdighetene i positiv retning. De estimerte leseferdighetene vil øke med nesten 5 poeng for hvert inntektshopp dersom de andre variablene holdes konstante. Tilsvarende observerer vi at de estimerte leseferdighetene øker med antall bøker i hjemmet og tidlige leseevner. Leseferdighetene synker derimot med økt innvandrerandel i klassen. Variabelen *par_edu* beskriver utdanningsnivået til foreldrene, der de høyest utdannede er i gruppe 1 og de lavest utdannede er i gruppe 5. Et hopp i kategorinivå fører i gjennomsnitt til en reduksjon i estimerte leseferdigheter på 10.91 poeng. Et fall i kategorinivå samsvarer med et hopp i utdanningsnivå, og det er altså en positiv sammenheng mellom nivået på utdanning og leseferdigheter.

6. Konklusjon

Resultatene fra de estimerte modellene tilsier at innvandrere har svakere leseprestasjoner enn øvrige elever, både i Norge og Tyskland. Det negative gapet viser seg å være svært robust når kontrollvariabler inkluderes i regresjonen. Vi har derimot ikke grunnlag for å si at dette gapet er ulikt mellom landene, som begge har ført relativt lik innvandringspolitikk.

Modellene viser at prestasjoner påvirkes av flere sosiodemografiske variabler som inntekt, kjønn og foreldres fødested. Jenter presterer vesentlig bedre enn gutter, noe som også er robust mot inkludering av kontrollvariabler. Kjønnsforskjellen er større i Norge enn i Tyskland, men vi har ikke grunnlag for å si at kjønnsforskjellen mellom innvandrere og ikke-innvandrere er ulik. Resultater fra regresjonsanalysen viser at foreldrenes inntektsnivå og fødested påvirker elevers leseferdigheter. Det viser seg at tyske elever gjør det bedre enn norske elever for et hvert inntektsnivå. Elever med foreldre i høye inntektsgrupper gjør det bedre enn elever med foreldre som har lav inntekt i begge land.

Til slutt ser vi at dersom vi inkluderer flere variabler i modellen, vil forklaringskraften stige. I vår første restriktive modell der vi kun har med interessevariablene, er forklaringskraften på 6.7 %. Vår siste modell har høyest forklaringskraft på 31.2%. Her har vi avdekket en rekke variabler som forklarer 31.2% av variasjonen til leseferdighetene.

7. Litteraturliste

Bakken, Anders. (2016). *Ulike perspektiver på skoleresultatene til barn og unge med innvandringsbakgrunn.* Tilgjengelig på: https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Ulike-perspektiver/

Datamaterialet fra PIRLS (2001) er tilgjengeliggjort på den interne NTNU-plattformen, *Blackboard*, av Institutt for Samfunnsøkonomi

Hanushek, Eric A. (2008). *Education Production Functions*. Tilgjengelig på: http://hanushek.stanford.edu/publications/education-production-functions (Hentet: 30. mars 2020)

Hanushek, Eric A. (2020). Education Production Functions. *The Economics of Education: A Comprehensive Overview, s. 161-170.* Tilgjengelig på: http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%202020%20Education%20Production%20Functions.pdf

Mathä, Thomas Y., Porpiglia, Alessandro & Sierminska, Eva (2011). *The Immigrant: Native Wealth gap in Germany, Italy & Luxembourg.* Tilgjengelig på: https://www.ecb.europa.eu//pub/pdf/scpwps/ecbwp1302.pdf

OECD (2012). *Language in a Global World; Learning for better Cultural Understanding.* Tilgjengelig på: https://www.oecd-ilibrary.org/education/languages-in-a-global-world_9789264123557-en (Hentet 06.05.20)

Pedersen, Jørn Wichne. (2018) *Skole og utdanning i Tyskland.* Tilgjengelig på: https://snl.no/Skole_og_utdanning_i_Tyskland (Åpnet: 13. mai 2020).

Raaum Oddbjørn, Stiftelsen Frischsenteret for samfunnsøkonomisk forskning. (2003). *Familiebakgrunn, oppvekstmiljø og utdanningskarrierer*. Tilgjengelig på: https://www.ssb.no/a/publikasjoner/pdf/sa60/kap-6.pdf (Hentet: 05.05.2020)

Rietig Victoria, Müller Andreas (2016). *The New Reality: Germany Adapts to Its role as a Major Migrant Magnet*. Tilgjengelig på: https://www.migrationpolicy.org/article/new-reality-germany-adapts-its-role-major-migrant-magnet (Hentet 10.05.2020)

Senter for leseforskning. (2001). *Slik leser 10-åringer i Norge.* Tilgjengelig på: https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/5/pirls_norsk_del_rapport.pdf

Statistisk sentralbyrå (2001). *Oppdaterte tall om innvandrere 2001*. Tilgjengelig på: https://www.ssb.no/befolkning/artikler-og-publikasjoner/oppdaterte-tall-om-innvandrere--42215 (Hentet: 13. mai 2020)

Statistisk sentralbyrå (2019). *Nasjonale prøver.* Tilgjengelig på: https://www.ssb.no/utdanning/statistikker/nasjprov/aar (Hentet 20.03.20)

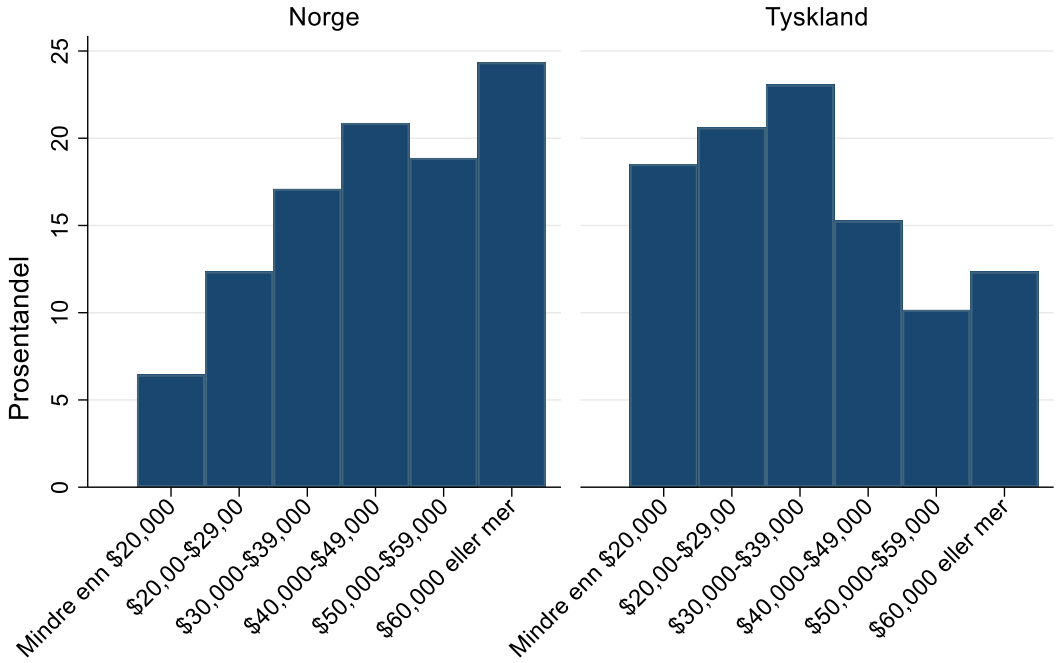
Thomas, R. (2005) *Using statistics in economics.* London: McGraw Hill.

UDIR. *PIRLS*. Tilgjengelig på: https://www.udir.no/tall-og-forskning/internasjonale-studier/pirls/ (Hentet: 9. mars 2020)

Vera Hanewinkel og Jochen Oltmer (2018) *Germany's Migration Policies.* Tilgjengelig på: https://www.bpb.de/gesellschaft/migration/laenderprofile/262811/germany-s-migration-policies (Hentet: 11. mai 2020)

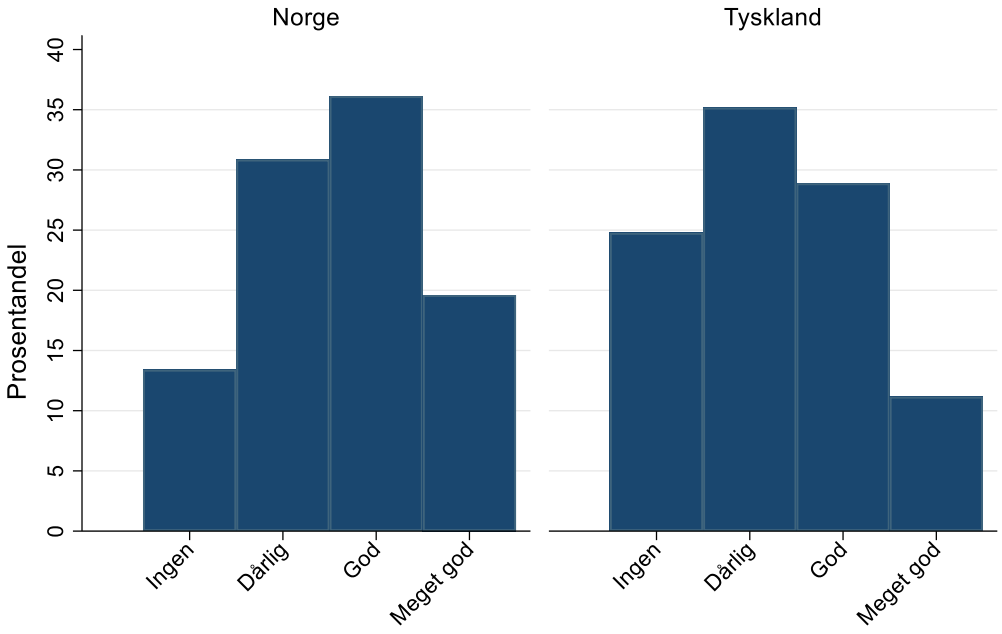
8. Appendiks

8.1.1. Husholdningenes årlige inntekt



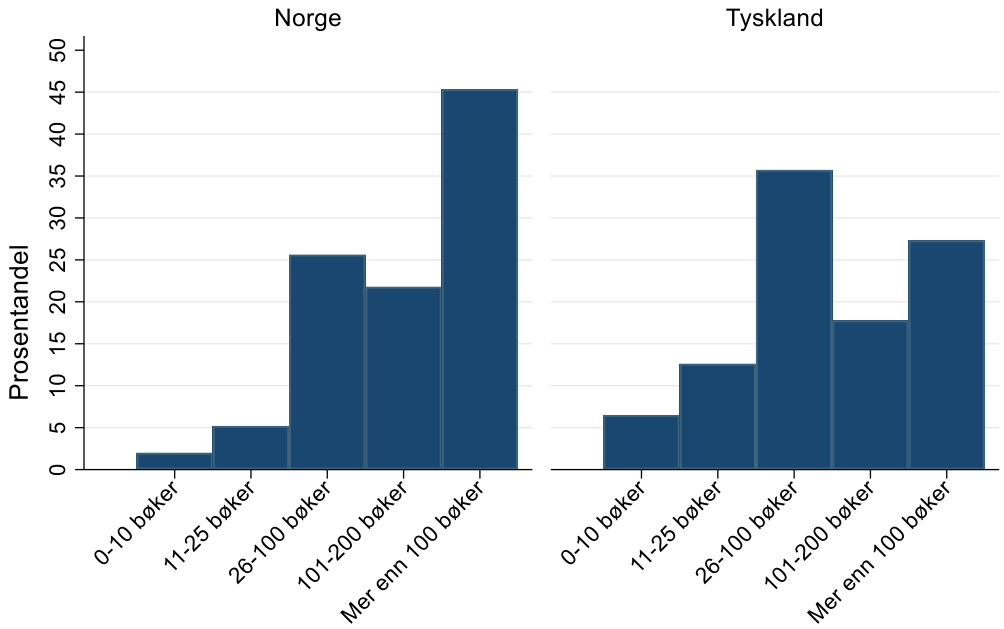
Variabel: *income*

8.1.2. Tidlige leseferdigheter blant elevene



Variabel: *early_ability*

8.1.3. Bøker i husholdningen



Variabel: *books_home*

8.1.4. Tabell – Oversikt over variabler

Variabelnavn	Beskrivelse	Variabeltype
read	Poengsum for leseferdigheter	Avhengig, kontinuerlig
par_not_born	Elevens foreldre ikke født i landet	Interessevariabel, dummy
girl	Kjønn	Interessevariabel, dummy
income	Husholdningens årlige inntekt målt i US$	Kontrollvariabel, kategori 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: $60,000 eller mer
books_home	Antall bøker i husholdningen	Kontrollvariabel, kategori 1: 0-10 bøker 2: 11-25 bøker 3: 26-100 bøker

		4: 101-200 bøker 5: Mer enn 200 bøker
early_ability	Tidlige leseferdigheter	Kontrollvariabel, kategori 1: Ingen 2: Svake 3: Gode 4: Meget gode
par_edu	Foreldrenes høyest fullførte utdanning	Kontrollvariabel, kategori 1: Universitet 2: Høyere utdanning (universitet ekskludert) 3: Videregående 4: Ungdomsskole 5: Ikke fullført ungdomsskole
pct_abroad	Prosentandel av elever født i et annet land på skolen	Kontrollvariabel, kategori 1: 0-10% 2: 11-25% 3: 26-50% 4: 51% eller mer
grmny	Eleven er tysk	Dummyvariabel

Inneholder utvalgte variabler fra datasettet

8.1.5. Tabell – Prosentandel av elevene på skolen som er innvandrere

	Norge	Tyskland
0 – 10 %	87.68 %	65.02 %
11 – 25 %	10.63 %	22.63 %
26 – 50 %	1.69 %	8.93 %
Mer enn 50 %	0 %	3.42 %

8.1.6. Tabell – Inntektsfordeling blant innvandrere og ikke-innvandrere i Norge og Tyskland

	Innvandrer i Norge	Født i Norge	Innvandrer i Tyskland	Født i Tyskland
Mindre enn $20,000	24.65 %	5.41 %	34.23%	15.22 %
$20,000 - $29,999	20.42 %	11.83 %	31.26 %	18.66 %
$30,000 - $39,999	14.08 %	17.10 %	17.40%	24.02%
$40,000 - $49,999	14.08 %	21.37 %	8.20 %	16.68 %
$50,000 - $59,999	10.56 %	19.47 %	4.10 %	11.50 %
$60,000 eller mer	16.20 %	24.81 %	4.81 %	13.91 %

8.1.7. Tabell – Korrelasjonsmatrise for variabler i modellene

	read	*par_no~n*	*girl*	*income*	*grmny*	*early_~y*	*books_~e*	*par_edu*	*pct_ab~d*
read	1.0000								
par_not_born	-0.1548	1.0000							
girl	0.1071	-0.0043	1.0000						
income	0.1865	-0.2247	0.0034	1.0000					
grmny	0.2994	0.1682	0.0085	-0.2517	1.0000				
early_abil~y	0.1443	0.0334	0.1299	0.0205	-0.1844	1.0000			
books_home	0.2633	-0.2737	-0.0008	0.4308	-0.2129	0.0273	1.0000		
par_edu	-0.2512	0.1502	0.0007	-0.4189	0.149	-0.0132	-0.4381	1.0000	
pct_abroad	-0.0179	0.3014	0.01	-0.1625	0.2853	-0.0106	-0.1622	0.1441	1.0000

8.1.8. Korrelasjonsmatrise for lærerkarakteristika på avhengig variabel

	read	*teacher_exp*	*teacher_cert*
read	1.0000		
teacher_exp	0.0449	1.0000	
teacher_cert	-0.0146	-0.1005	1.0000

8.2. Heteroskedastisitet

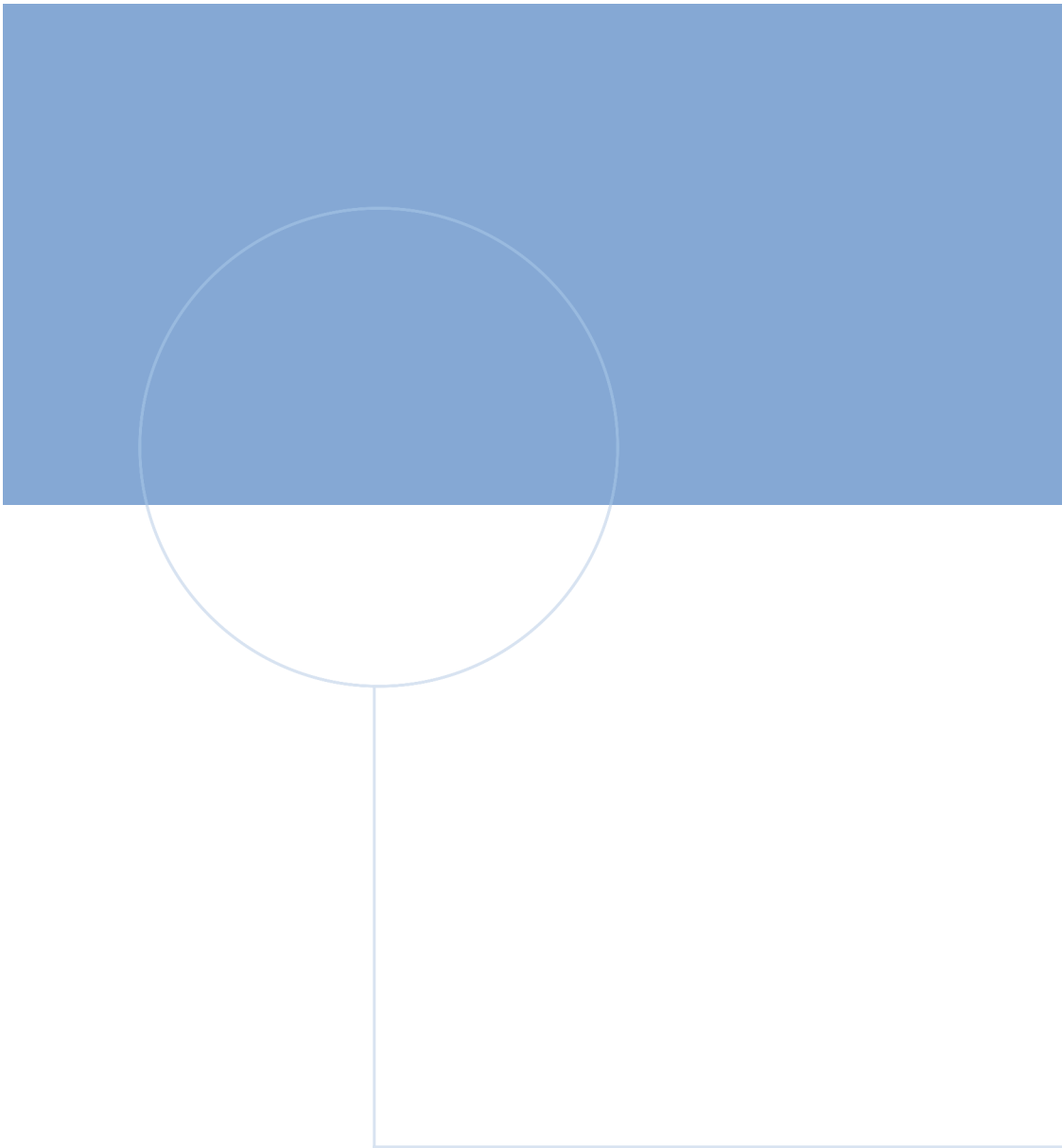
En av forutsetningene for OLS er at støyleddene har konstant varians, altså homoskedastistitet. Vi ønsker å undersøke om denne restriksjonen stemmer for å teste kvaliteten på modellen vår. Vi tester da for om restleddet har varierende varians, og dermed er heteroskedastisk. Ved heteroskedastisk støyledd er ikke lenger OLS-estimatoren effisient og det eksisterer en estimator med lavere varians. En konsekvens er at formelen for estimatorens standardavvik blir feil. For å korrigere for dette må det brukes såkalte robuste standardavvik (Thomas, 2005, s. 480).

Vi benytter Breusch-Pagan testen for å teste for heteroskedastisitet, da dette er mulig i *Stata*. Her undersøkes det om variansen i støyleddene fra en lineær regresjon er betinget av verdiene i de uavhengige variablene. Denne testen vil altså kontrollere om forutsetningen om at alle støyledd har konstant varians holder. I testen tar vi utgangspunkt i modell 5. Nullhypotesen vår sier at støyleddet er homoskedastisk, mens alternativhypotesen er at det er heteroskedastisisk.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of read

         chi2(1)      =   203.11
         Prob > chi2  =   0.0000
```

Testen viser at vi må forkaste nullhypotesen og konkludere med heteroskedastisitet. Det betyr altså at forklaringsvariablene vil påvirke variansen, og dermed også t-verdiene. Dette vil påvirke nivået for hvor vi kan forkaste hypoteser til et gitt signifikansnivå. Som konsekvens av heteroskedastisitet velger vi derfor å utføre regresjonene våre med robuste standardavvik.



NTNU

Kunnskap for en bedre verden