Bacheloroppgave

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

Kjernlie, Erik
Næss, Simen Emil Eide
Sundal, Marie Roholt

# Hvilke faktorer tilknyttet læreren påvirker elevenes leseferdigheter i Norden sammenlignet med land som tidligere tilhørte Sovjetunionen?

Bacheloroppgave i Samfunnsøkonomi
Veileder: Prof. Bjarne Strøm & Robin Valenta
Mai 2020

NTNU
Kunnskap for en bedre verden

Kjernlie, Erik
Næss, Simen Emil Eide
Sundal, Marie Roholt

# Hvilke faktorer tilknyttet læreren påvirker elevenes leseferdigheter i Norden sammenlignet med land som tidligere tilhørte Sovjetunionen?

Bacheloroppgave i Samfunnsøkonomi
Veileder: Prof. Bjarne Strøm & Robin Valenta
Mai 2020

Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
Institutt for samfunnsøkonomi

# Forord

Oppgaven som her presenteres inngår i faget SØK2901 - Bachelor i samfunnsøkonomi som en avsluttende oppgave for studieprogrammet Samfunnsøkonomi, bachelorprogram 3-årig. Hovedformålet med oppgaven er å kunne vise fortrolighet med metodeverktøy i statistikk og økonomi, samt tolke og forklare resultater fra økonometriske studier utfra en egen formulert problemstilling.

Programvaren som er benyttet for å løse oppgavens problemstillinger er det statistiske metodeverktøyet STATA som henter inn relevante data fra gitte datasett og som kan benyttes for å teste relevante hypoteser. Kodespråket Python har blitt brukt til å generere noen av figurene som presenterer datamaterialet i oppgaven, samt for søk etter mønstre og interne relasjoner i dataen som vil bli presentert i resultatene av oppgaven.

En egen formulert problemstilling og bakgrunn for oppgaven vil bli introdusert i det første kapittelet, etterfulgt av relevant teori, empiriske beregninger, resultater og tolkninger i de påfølgende kapitlene.

# Innhold


**1 Innledning** **1**

**2 Teoretisk rammeverk og tidligere litteratur** **2**
2.1 Skoleproduksjonsfunksjon . . . 2
2.2 Litteratur og tidligere studier . . . 2

**3 Empirisk strategi** **5**
3.1 Empirisk strategi - OLS-metoden . . . 5
3.2 Metodeforutsetninger og determinasjonskoeffisient . . . 6
3.3 Korrelasjon mellom variabler . . . 6
3.4 Hypotesetesting . . . 7

**4 Presentasjon av data og deskriptiv analyse** **8**
4.1 Informasjon om datasettet . . . 8
4.2 Valg av variable . . . 9
4.2.1 Valg av avhengig variabel . . . 10
4.2.2 Fremgangsmåte for valg av interesse- og kontrollvariabler . . . 10
4.2.3 Valg av interessevariabler . . . 12
4.2.4 Valg av kontrollvariabler . . . 15
4.3 Fordeler og begrensinger ved datasettet . . . 18

**5 Regresjonsanalyse** **20**
5.1 Lærerens påvirkning på leseferdighetene . . . 20
5.1.1 Regresjonsanalyse og hypotesetesting . . . 20
5.1.2 Empiriske resultater . . . 21
5.2 Lærerens påvirkning på leseferdighetene i Sovjetunionen sammenlignet med Norden . . . 23
5.2.1 Lineær regresjon og hypotestetesting . . . 23
5.2.2 Empiriske resultater . . . 24
5.2.3 Sammensetningen av læreregenskaper for lærere i Norden vs. Sovjet 26

**6 Sammendrag og konklusjon** **28**

**7 Appendix** **29**
7.1 Valg av variable . . . 29
7.1.1 VSelect . . . 29
7.1.2 Deskriptiv statistikk for kontrollvariabler . . . 30
7.2 VIF-test . . . 31
7.3 Chow-test . . . 32
7.4 Regresjonsanalyser . . . 33
7.5 Mønstergjenkjenning . . . 38

**Kapittel 1**

# Innledning

I samfunnsøkonomisk teori er det kjent at utdanning vil ha positive avkastninger, både privatøkonomiske og samfunnsøkonomiske. Den privatøkonomiske avkastningen gis uttrykk for i økt inntekt og økte nyttenivåer, ofte målt ved den prosentvise inntektsgevinsten som følger med ett år med mer utdanning. Sett fra et samfunnsøkonomisk perspektiv vil økt utdanning kunne medføre økt verdiskapning, gitt en optimering av bruk av arbeidsstyrkens kvalifikasjoner og evner (Barth, 2005). Om samfunnsøkonomisk eller privatøkonomisk avkastning er størst vil i tillegg ha betydning for hvilken utdanningspolitikk et land ønsker å føre, særlig med tanke på om arbeidsstyrkens kvalifikasjoner faktisk kommer nytte i landets produksjon, om individet har økonomiske ressurser eller støtteordninger til å ta utdanning og om tilleggseffekter på f.eks. kriminalitet kan synes (Barth, 2005).

Det er dog allment antatt at utdanning krever store offentlige investeringer. Dette både med tanke på fasiliteter og infrastruktur, ivaretagelse av solide fagmiljøer, gode arbeidsvilkår for lærere og lignende, men også i form av å kunne gi elever gratis utdanning med god utdanningskvalitet. Det er i dette tilfellet også interessant å undersøke om et land med begrensede ressurser per person kan tilby en utdanning av samme kvalitet som et land med store ressurser per person. Bakgrunnen og motivasjonen for denne oppgaven er å undersøke om elever fra land med mye ressurser viser høyere prestasjoner enn elever fra land med mindre ressurser. Et lands ressurser og velstand kan måles med ulike metoder, blant annet hjelp av deres bruttonasjonalprodukt. I denne oppgaven undersøkes det nærmere om elever fra land som tidligere har vært under et kommunistisk styre presterer forskjellig sammenlignet med elever fra nordiske land.

I denne oppgaven antas det at kvalitet på utdanning kan måles med elevprestasjoner i form av leseferdigheter. Dette er en antagelse med store usikkerhetsmomenter, men likevel en indikasjon på elevenes utbytte av utdanning gitt at dataene er sammenlignbare. Formålet med denne oppgaven er å gjennomføre empiriske studier for å avdekke avviket mellom elevprestasjoner i nordiske og øst-europeiske land, med fokus på lærernes påvirkning av hver enkelt elev. Det blir presentert en analyse som ser på sammenhengen mellom leseferdighetene til elevene og de ulike karakteristikkene tilknyttet lærerne for de ulike regionene. Problemstillingen i denne oppgaven lyder derfor som følger:

**Hvilke faktorer tilknyttet læreren påvirker leseferdighetene til en elev i Norden sammenlignet med en elev i eks-Sovjetunionen?**

For å avgjøre hvilke faktorer som differer seg for lærere i eks-Sovjetunionen sammenlignet med Norden undersøkes det først hvordan læreren generelt påvirker leseferdighetene til den enkelte elev. Følgende spørsmål stilles derfor for å belyse problemstillingen med et bedre perspektiv:

**Hvilke faktorer tilknyttet læreren påvirker leseferdighetene til en elev?**

For å besvare problemstillingen brukes det et datasett fra PIRLS-undersøkelsen; en internasjonal leseundersøkelse om leseferdigheter blant elever i fjerde klasse. Datasettet fra PIRLS undersøkelsen gjennomført i 2001 inneholder data for 35 land, med representanter fra både Norden og Øst-Europa. Fra Norden benyttes data for Norge, Island og Sverige, mens for de tidligere sovjetiske landene benyttes data for Latvia, Moldova og Russland.

**Kapittel 2**

# Teoretisk rammeverk og tidligere litteratur

Kapittel 2 presenterer det teoretiske rammeverket rundt skoleproduksjonsfunksjoner samt funn tidligere studier som er gjort knyttet til faktorer som påvirker elevpresentasjoner. Kapittelet tar også for seg en begrenset beskrivelse av de ulike landenes utdanningspolitikk.

## 2.1 Skoleproduksjonsfunksjon

I samfunnsøkonomisk teori produksjonsfunksjoner ofte benyttet for å optimere kvantum av innsatsfaktorer og det tilhørende nivået på en produksjon:

$$q = F(x_i, x_j..), i = \infty, j = \infty \tag{2.1}$$

Funksjonen beskriver hvordan produksjonen q avhenger av ulike innsatsfaktorer x. Ettersom valg (mengde og type) og tilhørende effekter av innsatsfaktorer kan varierer med tanke på tid er det ofte hensiktsmessig å skille mellom langsiktig og kortsiktig tidsperspektiv.

Oppgavens problemstilling kan også tilnærmes ved bruk av slike produksjonsfunksjoner, hvor forklaringsvariablene (lærernes ansiennitet og utdanningsnivå, klassestørrelse skolekarakteristikka, elevenes familiebakgrunn og husholdningsinntekt, statens involvering osv.) er innsatsfaktorer og elevpresentasjoner angis som produksjonen. Slike skoleproduksjonsfunksjoner (Education production functions) er en anerkjent tilnærming for å undersøke hvilke variabler som er utslagsgivende for evnenivået til elever (Hanushek, 2020). Funksjonene vil dessuten følge formen som beskrevet i uttrykk (2.1) med q som mål på elevprestasjon (output) og $x_i$, $x_j$ som ulike kontrollvariabler/forklaringsvariabler (input).

Det er også viktig å påpeke utdanning i seg selv er betraktet som en viktig faktor som har positive effekter på arbeidsmarket, husholdningers samlede inntekt samt et lands vekst og utvikling (Barth, 2005). Ved å sammenligne flere land er det dermed mulig å undersøke hvordan utdanningsnivå (direkte) og elevpresentasjoner (indirekte) kan bidra til vekst og økt velferdsnivå. For å oppnå en maksimal optimering av produksjonsfunksjoner og en påfølgende effektiv vekst vil det derimot være viktig å undersøke hvilke faktorer som bidrar til utdanning (med tanke på kvalitet og elevpresentasjoner), dvs. hvilke faktorer knyttet til hvordan kunnskap formidles og læres som bidrar til fremgang (Hanushek, 2020).

## 2.2 Litteratur og tidligere studier

Lese og skriveferdigheter er regnet som noen av de mest essensielle ferdighetene som må ligge til grunne for videre utdanning og læring i et samfunn. Landene som undersøkes i denne oppgaven er alle en del av eller delaktige i EU-områdets arbeid med utdanning som virkemiddel for bærekraftig og inkluderende vekst. Ofte vil leseferdigheter bli brukt som mål på hvordan et land ligger an i forhold til et annet. Slike sammenligninger er basert på tester og studier organisert av for eksempel IEA (International Association for the Evaluation of Educational Achievement). Den mest omfattende og sammenlignbare leseferdighetstesten som har blitt holdt i landene som inngår i oppgaven er den den

internasjonale leseundersøkelsen PIRLS (Progress in International Reading Literacy Study) som omfatter leseferdigheter for elever mellom 9 og 14. år. Den siste testen, utført i 2016 for 10-åringer, viser at Russland, Latvia, Norge og Sverige ligger alle godt over det internasjonale gjennomsnittet på 500 poeng. Alle land hadde også hatt en positiv utvikling, med gjennomsnittlige resultater fra 2016 tilsvarende 581 (Russland), 559 (Norge), 555 (Sverige) og 548 (Lithauen) (IEA, 2016a). Russland var også det landet som oppnådde høyest gjennomsnittlig poengscore sammenlignet med de rundt 50 landene som deltok. Blant 10-åringene som deltok skal er det rapportert at alle hadde hatt 4-års skolegang. Selv om Moldova og Island deltok ikke i undersøkelsen, er det grunn til å tro at resultatene ikke avviker særlig fra de andre landene selv på bakgrunn av lignende skolesystemer om dette dog må undersøkes videre.

De kanskje viktigste funnene fra 2016 undersøkelsen er hvilke faktorer som bidrar til positiv utvikling i leseferdigheter. IEA rapporten (IEA, 2016b) påpeker følgende faktorer og trender:

- Jenter tenderer til å oppnå høyere poengsum for leseferdigheter enn gutter. Dette er en trend basert på undersøkelser fra 2001 og frem mot 2016.
- Elever som har stabile hjemmeforhold tenderer til å ha høyere leseferdigheter. Positive faktorer som trekkes frem er; digitale verktøy tilgjengelig i hjemme, foreldres utdanning og leseferdigheter, foreldrenes holdninger til lesing og utdanning.
- Tidlig lesetrening. Faktorer som trekkes frem i rapporten er; alder ved skolestart og om foreldre har bidratt til tidlig lesetrening.
- Elever som tilhører skoler som har høyt fokus på utdanningskvalitet og har gode ressurser tenderer til å oppnå større poengsum for leseferdighet. Særlig fokus på lesetrening og oppfølging blir påpekt som en positiv faktor.
- Elever med gode leseferdigheter føler seg trygge når de er på skolen.
- Elever med gode leseferdigheter har lærere som er kvalifiserte.
- Elever som oppnår høy poengsum går ikke på skolen sulten eller trøtt over lengre tid.
- Elever med gode leseferdigheter rapporterer at de har en positiv holdning og motivasjon til lesing.

Funnene fra IEA (2016b) kan i stor grad synes å være sammenfallende med konklusjonene fra Hanushek (2020); det er ikke nødvendigvis en sammenheng mellom hvor store, relative økonomiske ressurser et land gir som input i en skoleproduksjonsfunksjon og elevers læringsutbytte (output). Det som er avgjørende er hvordan tilgjengelige ressurser allokeres. Hanushek (2020) påpeker at særlig lærerens kvalifikasjoner er en faktor som kan ha størst innvirkning på elevers læringsutbytte. Lærerens kvalifikasjon kan i en skoleproduksjonsfunksjon tolkes som en parameter som omfatter læringsvekst, og med en positiv sammenheng til elevers læringsutbytte (Hanushek, 2020). Studien (Hanushek, 2020) beskriver argumenterer også for at dersom en elev har hatt en god lærergjennom flere sammenhengende år vil dette ha en stor innvirkning på elevers prestasjoner tatt elevers fremtidige inntekter i betraktning. En viktig bemerkning til Hanuscheks argumenter er at det er noe utydelig hva som defineres som en god lærer", dvs. hvilke faktorer og undersøkelser som ligger til grunne for uttrykket (pedagogiske ferdigheter, egenmotivasjon og innsats, faglige ferdigheter, type og antall års erfaring osv.).

Når det kommer til spørsmålet om hvilken betydning lærertetthet og klassestørrelse har for elevenes læring gir ikke tidligere studier et entydig svar. Det er dog flere metodiske utfordringer med å sammenligne resultater fra forskning på klassestørrelse og lærertetthet, og synes avhenge av hvilket fagfelt (økonomi eller pedagogikk) forskningen springer utfra. Det er også ulike syn på om elevresultater alene kan gi en god og representativ beskrivelse av utdanningskvalitet og elevutbytte (Utdanningsforbundet, 2017). Enkelte studier viser derimot at det ikke nødvendigvis er en sammenheng mellom elevpresentasjoner og lærertetthet (Jelstad, 2015, Hanushek 2020), men at det først og fremst er ressursinnsats (kvalitet, kvantitet) og skolenes optimering av ressurser som vil være avgjørende for hvordan gode elevpresentasjoner opprettholdes (Bonesrønning, 2008).

Hvordan ressurser er allokert og hvilke faktorer som favoriseres i de ulike landene er synes derimot vanskelig å finne gode og objektive kilder til. Det er også motstridende informasjon relatert til hvor mye de ulike landene faktisk investerer i utdanningssektoren og hva investeringene faktisk innebærer. En generell tilnærming til BNP/innbygger er det derfor valgt som en indikasjon for hvor store økonomiske ressurser et land har til å tilby utdanning til hvert individ. Ser man da videre på de landene som er inkludert i oppgaven ser man at Norge (450 milliarder US dollar/ 5,42 millioner innbyggere) er det landet som har høyest BNP per innbygger etterfulgt av Island (24,7 milliarder US dollar / 0,34 millioner innbyggere), Sverige (575 milliarder US dollar/ 10,09 millioner innbyggere), Lithauen (54,3 milliarder US dollar/ 2,72 millioner innbyggere), Russland (1750 milliarder US dollar/ 143,93 millioner innbyggere). Verdiene er basert på GDP oversikter fra TradingEconomics (2020) for hvert av landene og populasjonstall fra Worldometers (2020). Tallene er basert på statistikk fra 2019.

Videre er det allment kjent at land som tidligere tilhørte Sovjetunionen opplevde økonomiske nedgangstider etter unionsoppløsningen. Tilnærmingen til en effektiv markedsøkonomi ble langvarig og vanskelig i land som Moldova (FN, 2020a), mens Russland styrket seg sterkt økonomisk på begynnelsen av 2000-tallet da olje og gass eksport bidro til økte statsinntekter. Latvia på sin side har den dag i dag store fattigdomsutfordringer, selv om landet de siste årene har opplevd økonomisk fremgang (FN, 2020b). Felles for tidligere sovjetstater er at rikdommene er samlet på få hender, noe bidrar til store forskjeller i levestandard og inntekter (Carlsen, 2015). For enkeltindivider uten sterke økonomiske ressurser kan det da oppstå sterke insentiver til å velge tidlig inntekt fremfor utdanning, noe som i generelle tilfeller kan redusere elevers faglig støtte i hjemmet (Regjeringen, 2019). De nordiske landene som har et økonomisk styresett rettet mot en blandingsøkonomi har vist stor avkastning for den enkelte innbygger (materielt og i form av statsinvesteringer), og andelen av befolkningen som har fullført høyere utdanning ligger på topp i Europa (Gornitzka, 2003).

**Kapittel 3**

# Empirisk strategi

Kapittel 3 omfatter en beskrivelse av det teoretiske rammeverket som er benyttet for å løse oppgavens problemstillinger. Kapittelet tar for seg OLS-metoden og relevante hypoteseteser på en generell basis som grunnlag for videre analyse i kapittel 4 og 5.

## 3.1 Empirisk strategi - OLS-metoden

Minste kvadraters metode, her referert til som OLS-metoden, er en matematisk regresjonsmetode ofte brukt i statistiske elle økonometriske sammenhenger for å etablere årsakssammenheng mellom variabler . Disse sammenhengene kan så benyttes for å estimere observerbare eller ukjente variabler i et gitt datasett, samt forklare forholdet mellom en endogen variabel og en eller flere eksogene variabler (Thomas, 2005). En viktig forutsetning for bruk av OLS-metoden er at årsakssammenhengene mellom variablene er lineære. Variablene som benyttes er henholdsvis en endogen responsvariabel Y som uttrykker det totale utfallet, samt eksogene forklaringsvariabler Xi. En forenklet sammenheng mellom variablene Y og X kan uttrykkes som:

‘

$$y = \alpha + \beta x \tag{3.1}$$

Dersom sammenhengen (Ligning 1) illustreres grafisk vil man se at den former en lineær linje hvor konstanten (alpha) viser hvor linjen krysser y-aksen, mens konstanten (beta) er den lineære linjens stigningstall. Hensikten med regresjon er da å finne den beste mulige tilpasningen til linjen ved å bruke OLS-metoden. OLS-metoden minimerer variansen til Y, dvs. kvadratet kvadratet til avviket mellom den observerte og den estimerte verdien. Dette gir så estimater til (alpha) og (beta). Den beste tilpasningen til linjen, dvs. regresjonslinjen, kan da uttrykkes som:

‘

$$\hat{y} = \hat{\alpha} + \hat{\beta} x \tag{3.2}$$

hvilket kan omskrives til uttrykket:

$$Y_i = \alpha + \sum_{i=1}^{n} \beta_i x_i + \epsilon \tag{3.3}$$

Oppgavens hovedinteresse er å estimere parameteren foran lærer-koeffisientene ($\beta_i$) i en lineær regresjonsmodell, med og uten kontrollvariabler. Nærmere analyse vil bli presentert i kapittel 4 og 5.

## 3.2 Metodeforutsetninger og determinasjonskoeffisient

En viktig bemerkning er at den faktiske verdien på Y ikke nødvendigvis vil være lik forventningsverdien, E(Y). For å korrigere for dette avviket introduserer man støyleddet ($\epsilon$), som fanger opp andre og ofte ikke-målbare faktorer som kan påvirke både Y og den forventede verdien til populasjonen. For at små residualer ($\epsilon$) ikke skal kunne gi misvisende resultater som følge av endringer i den endogene variabelen, krever metoden at følgende forutsetninger er oppfylt (Thomas, 2005):

1. $E(\epsilon_i)$=0. Støyleddet ($\epsilon$) må ha en forventningsverdi lik null slik at avviket mellom estimerte og observerte verdier er lik null.
2. $Var(\epsilon_i) = \sigma^2 < \infty$. Variansen til støyleddet er konstant og uendelig for alle verdier.
3. $Cov(\epsilon_i, \epsilon_j) = 0$. Støyleddene er uavhengige av hverandre.
4. $\epsilon_i$ $N(0, \sigma^2)$. Støyleddet er normalfordelt slik at hypotesetesting kan utføres.

Til tross for at OLS-estimatene skal gi den beste regresjonslinjen gitt et bestemt datasett, er det ofte hensiktsmessig å beregne hvor godt linjen faktisk beskriver datasettet. Dette kan undersøkes ved bruk av determinasjonskoeffisienten $R^2$ (Thomas, 2005, s. 273). $R^2$ beskriver i hvor stor grad variasjoner i X kan gi utslag til variasjon i Y, dvs. forklaringsvariasjon (SSE) i forhold til total beregnet variasjon (SST) (Thomas, 2005). $R^2$ kan da uttrykkes som:

$$R^2 = \frac{SSE}{SST} = \frac{b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, 0 \leq R \leq 1 \tag{3.4}$$

Som man kan se fra ligning (3.4) vil $R^2$ være et konstant tall mellom 0 og 1 som øker desto større SSE er. Størrelsen på SSE kan forklares utfra hvor store variasjoner som finnes i ligningssettet og antall varianter av X som modellen inneholder. Det er dog viktig å påpeke at $R^2$ kan gi misvisende estimater, og at det derfor er mer hensiktsmessig å benytte en versjon av $R^2$ som tar hensyn til at flere variabler inkluderes i modellen (Thomas, 2005), som for eksempel Adjusted $R^2$ (MathWorks, 2020).

## 3.3 Korrelasjon mellom variabler

For å undersøke hvordan kvantitative variabler avhenger av hverandre (retning og styrke) vil man i statistikk og sannsynlighetsregning benytte korrelasjonskoeffisienten r. Hvordan populasjonsvariabler i et utvalg avhenger av hverandre kan da måles ved bruk av følgende ligning:

$$r = \frac{b^2 \sum (X - \bar{X})(Y - \bar{Y})}{\sum \sqrt{(X - \bar{X})^2}\sqrt{(Y - \bar{Y})^2}}, -1 \leq R \leq 1 \tag{3.5}$$

I tilfeller hvor det ikke er en lineær sammenheng mellom populasjonsvariable vil r=0, og alle andre verdier mellom -1 og 1 dersom det en lineær sammenheng finnes. Som beskrevet av Thomas (2005) er det ofte viktig å stille kritiske spørsmål til korrelasjonsresultatenes gyldighet. Korrelasjon er i all hovedsak definert som samvariasjon, og ikke nødvendigvis en eksakt sammenheng. I enkelte tilfeller kan man oppleve at en tredje variabel Z, som

kan påvirke Y og X estimater, medfører en årsakssammenheng mellom X og Y som ikke nødvendigvis er korrekt. I tillegg er det viktig å ha i mente at OLS-metoden i all hovedsak benyttes for å estimere individuelle påvirkninger av forklaringsvariablene X. En perfekt sammenheng vil i så tilfelle bryte med OLS-forutsetningene da det er kun en estimert kombinasjon av variablene som kan undersøkes (Thomas, 2005).

## 3.4 Hypotesetesting

For å undersøke om en statistisk antagelse eller påstand om egenskaper for en eller flere populasjoner holder benytter man hypotesetesting. Man presenterer da en nullhypotese ($H_0$) og en alternativhypotese ($H_1$). Oftest vil nullhypotesen presenteres som en antitese hvor man antar det ikke finnes en sammenheng eller korrelasjon mellom variablene. Alternativhypotesen derimot antar at det er en sammenheng, korrelasjon, mellom variablene og uttrykkes som en verdi ulik null. En tosidig hypotesetest kan da generelt uttrykkes som:

$$H_0 : \beta_i = 0, H_1 : \beta_i \neq 0 \tag{3.6}$$

Ettersom man ikke kan 100% kan avkrefte eller bekrefte en påstand, vil man i all hovedsak forsøke å teste nullhypotesen slik at dersom denne avkreftes vil det gi større insentiver til å bekrefte alternativhypotesen. For å kunne avgjøre om resultatene fra hypotesetestingen er statistisk signifikante, altså hvor stor sannsynligheten er for å måtte forkaste en nullhypotese, velges et signifikansnivå, $\alpha$. Ofte settes $\alpha = 0,05$ (%5 med et tilhørende konfidensintervall 95%. Konfidensintervallet på sin side beskriver hvor gode estimatene faktisk er. For problemstillingene presentert i denne oppgaven er signifikansnivå 0.05 benyttet.

Videre er det hensiktsmessig å bemerke at det finnes flere typer hypotesetester. I denne oppgaven fokuseres det på t-test og f-test. T-testen vil være symmetrisk rundt null og støtter en normalfordelingsform. Slike tester benyttes dersom man ønsker å undersøke om en spesifikk variabel bør tas med i regresjonsmodellen som etableres og gir svar på om en eventuell endring i modellens forklaringskraft er signifikant (Thomas, 2005). F-testen benyttes derimot i variasjonsanalyse hvor man ønsker å finne ut forklart og uforklart varians når man endrer regresjonsmodellen og ser på hvordan resultater endres når enkelte variabler legges til eller fjernes.

For å teste for strukturelle endringer i datasettet knyttet til om lærerens egenskaper påvirker resultatene til regresjonsmodellen forskjellig avhengig av land, benyttes Chow-testen (Gould, 2014). Ved å se om koeffisientene endres for en modell av en referansegruppe til sammenlikning med gruppen den skal testes mot kan man formulere en nullhypotese om at de er strukturelt like. Kan nullhypotesen avvises, og de to modellene viser seg å være ulike, kan man si at det er en strukturell ulikhet mellom modellene.

**Kapittel 4**

# Presentasjon av data og deskriptiv analyse

Dette kapitellet omfatter preprosessering av informasjonen som brukes videre i regresjonsanalysen. Kapitellet tar for seg deskriptiv analyse og seleksjon av variabler som ansees som relevante for å forstå hvordan lærere påvirker elevenes resultater.

## 4.1 Informasjon om datasettet

Denne rapporten fokuserer på hvordan læreren påvirker elevenes leseferdigheter i Norden sammenlignet med tidligere sovjetiske land. Studiet i denne rapporten er derfor avgrenset til datapunktene knyttet til Norge, Island, Sverige, Latvia, Russland og Moldova. Tabell 4.1 lister opp relevante variabler for analysen. Alle interessevariablene er faktorer tilknyttet lærerne i undersøkelsen, da fokuset i problemstillingen er hvordan læreren påvirker elevenes ferdigheter.

| **Variabel** | **Forklaring** | **Forkortelse** |
|---|---|---|
| Leseferdigheter | Poengsum i lesing | read |
| Kjønn* | Lik 1 hvis læreren er kvinne | teacher_fem |
| Alder* | Seks kategorier for lærerens alder:<br>1 - mindre enn 25 år<br>2 - 25 til 29 år<br>3 - 30 til 39 år<br>4 - 40 til 49 år<br>5 - 50 til 59 år<br>6 - minst 60 år | teacher_age |
| Sertifisering* | Lik 1 hvis læreren er sertifisert | teacher_certificate |
| Universitetsutdanning* | 1 hvis universitetsgrad, 0 hvis ikke. | teacher_edu |
| Erfaring* | Antall år med erfaring | teacher_exp |
| Samme lærer under et år | Lik 1 hvis sann. | sameteacher_1less |
| Samme lærer mer enn fire år | 1 hvis sann. | sameteacher_4plus |
| Kontrollvariabler | Inkluderer flere variabler som antas å påvirke leseferdighetene | $X_i$ |

Tabell 4.1: Forklaring av relevante interessevariabler med en forkortelse for hver variabel som er brukt i regresjonsanalyser. Variabler markert med stjerne (*) er verdier for læreren. Dette vil si at f.eks. Kjønn*er lærerens kjønn.

Datasettet består av 14.334 observasjoner for de nordiske landene og 10.645 observasjoner for land som tidligere tilhørte Sovjetunionen. For å lettere kunne sammenligne de eks-sovjetiske landene med landene i Norden, så er det hensiktsmessig å samle dataen for de nordiske og de eks-sovjetiske landene i samme datasett. Norden består da av Norge, Sverige og Island, mens Sovjetunionen inneholder datapunktene for Latvia, Russland og Moldova. Selv om det er forskjeller i leseprestasjoner innad i landene i hver av regionene, så er det trendene relatert til lærerkarakteristikken innad i hver av regionene det er aktuelt å undersøke i denne rapporten.

Tabell 4.2 viser deskriptiv statistikk for leseprestasjoner i Norge, Island, Sverige og Norden. I Norden utmerker Sverige seg med en høyt gjennomsnittlig lesescore på omtrent 565 og et lavere standardavvik enn de andre landene i Norden. Tabell 4.3 viser deskriptiv statistikk for leseprestasjoner for Latvia, Russland og Moldova, samt for et datasett som inkluderer alle de tre sovjetiske landene, hvilket som ofte refereres til som Sovjet eller Sovjetunionen i resten av oppgaven. Gjennomsnittlig leseprestasjon av alle elever i de eks-sovjetiske landene er omtrent samme som i Norden, med en gjennomsnittlig poengsum på ca. 534. Latvia har det høyeste gjennomsnittet, mens Moldova har det laveste. De øst-europeiske landene har omtrent like mange datapunkter for hvert land, mens Sverige er overrepresentert i den nordiske regionen med over halvparten av datapunktene.

| | Norden | Norge | Island | Sverige |
|---|---|---|---|---|
| Antall observasjoner | 14334 | 3459 | 3676 | 7199 |
| Gjennomsnitt | 535.36 | 498.26 | 512.79 | 564.70 |
| Standardavvik | 74.54 | 78.37 | 70.95 | 61.31 |
| Minimum | 228.06 | 228.06 | 252.66 | 318.68 |
| Maximum | 737.33 | 695.87 | 705.82 | 737.33 |

Tabell 4.2: Deskriptiv statistikk med antall observasjoner, gjennomsnitt, standardavvik, minimum og maksimum for leseferdighetene i Norge, Island, Sverige og Norden.

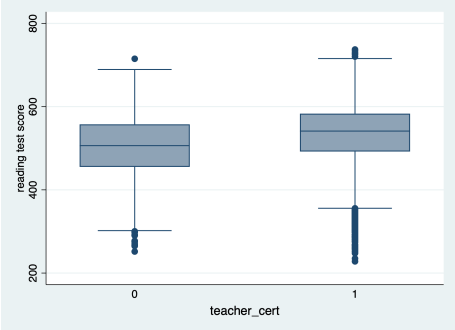
| | Sovjet | Latvia | Russland | Moldova |
|---|---|---|---|---|
| Antall observasjoner | 10645 | 3019 | 4093 | 3533 |
| Gjennomsnitt | 533.90 | 549.6144 | 529.99 | 492.78 |
| Standardavvik | 90.48 | 57.50 | 62.33 | 70.96 |
| Minimum | 428 | 332.25 | 252.08 | 247.60 |
| Maximum | 643 | 704.47 | 688.45 | 715.00 |

Tabell 4.3: Deskriptiv statistikk med antall observasjoner, gjennomsnitt, standardavvik, minimum og maksimum for leseferdighetene i Latvia, Russland, Moldova og Sovjet.

## 4.2 Valg av variable

Regresjonsanalyser benyttes for å beskrive sammenhengen mellom en eller flere uavhengige variabler og en avhengig variabel. For en lineærregresjon, så er sammenhengen mellom de uavhengige variablene og den avhengige variabelen beskrevet ved hjelp av en rett linje. Når en analyse gjennomføres med flere uavhengige variabler kalles det en multippel regresjonsanalyse. De uavhengige variablene det er interessant å undersøke i analysen kalles interessevariabler. Kontrollvariabler inkluderes i analysen for å utelukke at sammenhengen mellom den avhengige og de uavhengige variablene ikke skyldes tredje-variabler som ikke er tatt med i analysen. Disse holdes konstant gjennom analysene for å undersøke effekten av interessevariablene på den avhengige variabelen. Ved å legge til kontrollvariabler kan man utelukke at den påviste sammenhengen ikke skyldes disse egenskapene, da det er antatt at disse faktorene kan påvirke både den avhengige og uavhengige variablene. Hvis en variabel med forklarende kraft som er korrelerte med en eller flere av de andre uavhengige variablene er utelatt fra analysen, så kan det føre til et utelatt variabel problem.

Det er mange ulike variabler det kan være interessant å undersøke effekten av, og for regresjonsanalyser må man dermed bestille hvilke variabler som skal inngå i analysen. En interessant sammenheng mellom avhengig variabel og forklaringsvariabel kan eksempelvis være at man mistenker at en lærer sin sertifisering kan påvirke en elev sine leseferdigheter, slik man får et inntrykk av fra fordelingen til *read* over *teacher_cert* i figur 4.1. Figuren viser at elever som har en lærer som er sertifisert har en median med en høyere verdi for leseferdigheter og et mindre standardavvik. Merk at figuren ikke viser hvor mange som er sertifisert, da dette ikke er relevant for tolkningen.



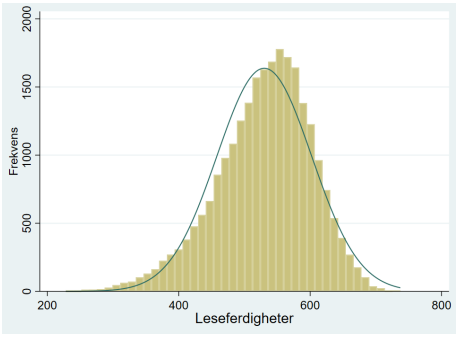

Figur 4.1: Fordelingen av elevers leseferdigheter ut i fra lærerens sertifisering

### 4.2.1 Valg av avhengig variabel

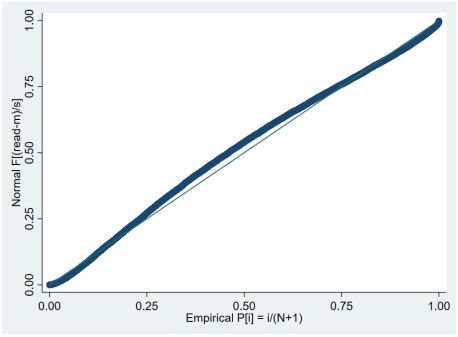
Et utgangspunkt som er helt prekært for å se på sammenhengene mellom leseferdigheter og de andre forklaringsvariablene i skoleproduksjonsfunksjonen er at leseferdighetene, gitt som variabelen *read*, må være normalfordelt og kontinuerlig. Om dette ikke var tilfellet ville man fått en ubalansert modell, noe som blir ofte referert til som *bias* i statistikken. Man kan se utfra figur 4.2 at leseferdighetene følger en normalfordeling. I plottet til venstre er y-aksen antall elever med en viss leseferdighet. Medianen til leseferdighet-variabelen er ca. 537, mens gjennomsnittet er rundt 530. Siden gjennomsnittet er litt lavere enn medianen, så er variabelen venstreskjev. Dette er et mål på at variabelen ikke er symmetrisk. Dette bekreftes av P-P plottet i samme figur. P-P plott blir i statistikken brukt for å sjekke om et datasett er normalfordelt. Hvis variabelen er normalfordelt, så vil punktene være fordelt på en rett linje. Det observeres at det er et lite avvik fra normalfordelingen, noe som bekrefter at det er en liten skjevhet i dataen. Skjevheten er dog ikke utslagsgivende for analysen og det anses videre ikke som et problem for regresjonsanalysen at det er en liten skjevhet i dataen. Variabelen *read* velges dermed som den avhengige variabelen, da hensikten av regresjonsmodellen er å lage en tilnærming av de forskjellige faktorene som påvirker en elev sine leseferdigheter.

### 4.2.2 Fremgangsmåte for valg av interesse- og kontrollvariabler

For å kartlegge de interne relasjonene i datasettet og få forståelse av hvor mange variabler som har høy forklaringskraft ble det benyttet innebygde variabelseleksjonsmoduler i Stata. For å få en pekepinn på hvilke variabler som burde bli brukt for en optimal lineær regresjonsmodell, så ble VSelect-modulen i STATA brukt. Denne modulen hjelper med å finne ut hvilke variabler man bør bruke i en lineærregresjon basert på Furnival and




(a) Fordelingen til leseferdighetene i et felles datasett for Norden og det tidligere Sovjetunionen.




(b) P-P plot av leseferdighetene. Perfekt normalfordeling er gitt ved den tynne, rette linjen. Det observeres et lite avvik mellom dataen og en perfekt normalfordeling.

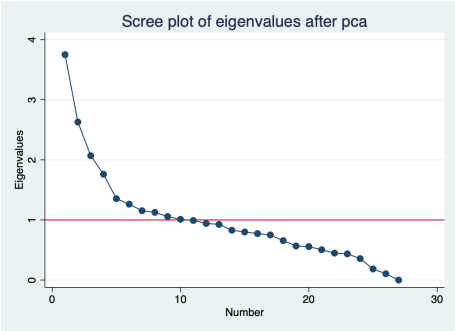
Figur 4.2: Egenskaper av variabelen tilknyttet leseferdighetene til en elevene

Wilsons leaps-and-bounds algoritme (Furnival and Wilson, 1974). En analyse med alle variablene gir resultatene vist i avsnitt 7.1.1 i appendix. Denne analysen beholder 24 av 27 variabler i datasettet, noe som tyder på at det er mange variabler som er viktige for regresjonsmodellen. Det er dog for mange variabler for en enkel regresjonsmodell.

For å teste forskjellige lineærtransformasjoner av kombinasjoner av variablene som brukes til å forklare sammenhenger internt i datasettet brukes Principal Component Analysis (PCA)(Wold, Esbensen, Geladi, 1987). Ønsket med denne metoden er å avdekke hvordan undergrupper av ulike kombinasjoner av variabler vil kunne utgjøre et datasett egnet til å beskrive relasjonene i dataen og dermed brukes i lineærregresjonen. En PCA tar utgangspunkt i variansen og korrelasjonen mellom variablene i datasettet. En PCA-analyse av datasettet brukt i denne rapporten viser at det trengs over 20 undergrupper for å redegjøre for variasjonen i den originale dataen, som vist i figur 4.3. Resultatet forteller at løsningsrommet i dataen er stort, og at mange av variablene trengs for å kunne forklare leseferdighetene til en elev.

For å undersøke variablenes grad av korrelasjon, så ble forklaringsvariablene undersøkt ved hjelp av VIF-tester og korrelasjonsmatriser. Dette er nyttig for å se på graden av den lineære sammenhengen mellom flere forklaringsvariabler i en multippel regresjonsmodell for å utelukke multikollinearitet. En VIF-test er en indikator som angir graden av multikollinearitet. Resultatet av en VIF-test gjennomført på leseferdighetene med alle variablene i datasettet er vist i figur 7.4 i appendix. En tommelfingerregel er at en poengsum på 1 eller mindre viser at variablene ikke er korrelert, mellom 1 og 5 tilsier at variablene er moderat korrelert, mens over 5 sier at variablene er høyt korrelert. Fra datasettet ser det dermed kun ut som elevens alder, fødselsår og fødselsmåned er korrelerte. Som regel vises graden av linearitet i en korrelasjonsmatrise for å se korrelasjonen mellom mange variabler. Dette er undersøkt nøyere i kapittel 4.2.4.

Både VSelect og PCA tyder på at det er mange variabler som bør inngå i en regresjonsanalyse for å få en best mulig modell. Disse verktøyene er kun benyttet for å gi et grunnleggende innblikk av hvilke variabler som er signifikante, og det tas ingen valg utelukkende basert på disse modellene. Det vektlegges derimot en økonomisk tolkning av hver variabel for å bestemme hvilke variabler som skal benyttes. Det er også viktig å understreke at hensikten




Figur 4.3: Eigenvalue til covariansmatrisene til hver container

med regresjonsmodellene er å undersøke påvirkningen av faktorene relatert til læreren på en elev, og ikke en optimal modell for å kalkulere leseferdigheter. De nevnte verktøyene gir dog en god veiledning på kvaliteten av modellene som benyttes.

### 4.2.3 Valg av interessevariabler

Problemstillingen i denne oppgaven er relatert til lærerens påvirkning av elevenes leseferdigheter. Det er derfor nødvendig å se på faktorene relatert til læreren. Dette er faktorer som lærerens kjønn, alder, erfaring og om han eller hun har et sertifikat. Det er også relevant å se på hvor lenge en elev har en bestemt lærer og om dette påvirker leseferdighetene. Alle disse variablene er tatt med som interessevariabler. Lærerens utdanning, *teacher_edu*, fjernes som interessevariabel, da det ikke finnes noen datapunkter for denne variabelen.

#### Deskriptiv statistikk for interessevariabler

Deskriptiv statistikk for interessevariabler er vist i tabell 4.4 og tabell 4.5. Det er verdt å merke seg at flere lærere er kvinner i det eks-sovjetiske landet sammenlignet med Norden, med en forskjell på ca. 10 prosentpoeng. For de øst-europeiske landene har kun 77% av lærerne et sertifikat, sammenlignet med 94% i Norden. Moldova utmerker seg med en andel på kun 42%, noe som bidrar til å dra ned snittet drastisk for Sovjetunionen. Når det gjelder antall elever som har hatt samme lærer i 4 år eller mer i det eks-sovjetiske landet, så gjelder dette over 66% av elevene. Dette er langt høyere enn i Norden, hvor kun 14% av elevene har hatt samme lærer i 4 år eller mer. Lærerens alder samt antall elever som har hatt samme lærer i et år eller mindre er relativt likt for begge regionene.

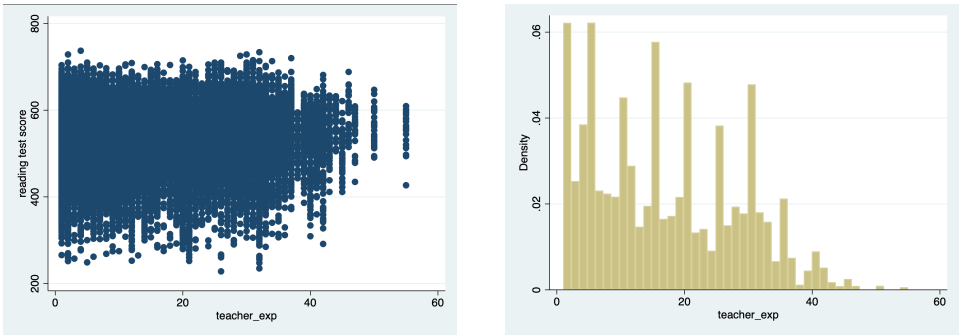
| | Norden | Norge | Island | Sverige |
|---|---|---|---|---|
| teacher_fem | 0.86 (0.35) | 0.87 (0.34) | 0.93 (0.25) | 0.81 (0.39) |
| teacher_age | 3.79 (1.18) | 3.94 (1.20) | 3.62 (1.09) | 3.80 (1.21) |
| teacher_cert | 0.94 (0.24) | 0.97 (0.17) | 0.91 (0.29) | 0.93 (0.25) |
| teacher_exp | 15.09 (11.15) | 16.51 (11.09) | 13.02 (10.34) | 15.40 (11.41) |
| sameteacher_1less | 0.01 (0.10) | 0 (0) | 0 (0) | 0.02 (0.13) |
| sameteacher_4plus | 0.14 (0.35) | 0.40 (0.49) | 0.10 (0.30) | 0.04 (0.18) |

Tabell 4.4: Deskriptiv statistikk med gjennomsnitt (og standardavvik i parentes) for variabler tilknyttet lærerkarakteristikken for de nordiske landene.

| | Sovjetunionen | Latvia | Russland | Moldova |
|---|---|---|---|---|
| teacher_fem | 0.96 (0.19) | 0.98 (0.15) | 0.99 (0.08) | 0.91 (0.29) |
| teacher_age | 3.55 (1.22) | 3.70 (1.22) | 3.63 (1.17) | 3.31 (1.24) |
| teacher_cert | 0.77 (0.42) | 0.91 (0.29) | 0.97 (0.18) | 0.42 (0.49) |
| teacher_exp | 19.58 (10.87) | 19.70 (11.35) | 20.35 (10.52) | 18.52 (10.79) |
| sameteacher_1less | 0.01 (0.10) | 0.02 (0.14) | 0 (0) | 0.02 (0.13) |
| sameteacher_4plus | 0.66 (0.47) | 0.76 (0.43) | 0.39 (0.49) | 0.90 (0.30) |

Tabell 4.5: Deskriptiv statistikk med gjennomsnitt (og standardavvik i parentes) for variabler tilknyttet lærerkarakteristikken for de land som tidligere var en del av Sovjetunionen.

Lærerens erfaring, *teacher_exp*, er den eneste kontinuerlige interessevariabelen. Et spredningsplott av denne variabelen er vist i figur 4.4. Dette plottet viser fordelingen av leseferdigheter sammen med antall års erfaring til læreren for hele datasettet. Til tross for at det er vanskelig å si noe om relasjonen, da mesteparten av datapunktene overlapper, så ser man at variansen reduseres med økt erfaring. Grunnen til dette kan blant annet være som følge av et mindre datagrunnlag for lærere med mange års erfaring, noe man kan se utifra histogrammet i samme figur. Det er naturlig nok bare et fåtall antall lærere som vil ha erfaring over 40 år grunnet høy alder, men det observeres manglende verdier for lærere med 48, 49, 51, 52, 53 og 54 års erfaring.

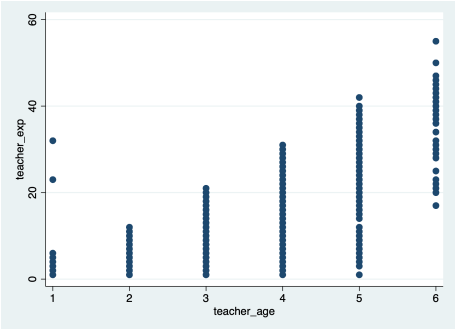



(a) Fordelingen av leseferdigheter over lærers erfaring

(b) Histogram for lærererfaring

Figur 4.4: Egenskaper om variabelen til lærerens erfaring og samspill med leseferdigheter

Allerede før korrelasjonskoeffisientene mellom lærernes alder og erfaring er undersøkt, vil det være naturlig å påpeke relasjonen deres. Lærerens alder, *teacher_age*, er en kategorisk variabel bestående av aldersgrupper. Når erfaringen til en lærer øker, er det også rimelig å anta at alderen øker. Et spennende fenomen ved land som har opplevd stor arbeidsledighet i fortiden er hvordan forholdet mellom de to variablene kan endres. Etterhvert som arbeidsledigheten går opp, vil det være naturlig å trekke mot sikre yrker, slik som læreryrket. Dette vil kunne resultere i flere lærere med høy alder og lite erfaring, og dermed en interessant forskjell mellom Norden og Sovjet, da de Østeuropeiske landene har opplevd en økonomisk resesjon etter oppløsningen av Sovjetunionen. Det er viktig å merke seg, som poengtert i avsnitt 3.3, at det til tross for en samvariasjon mellom lærerens erfaring og alder, vil variablene kunne ha forskjellig forklaringseffekt for den avhengige variabelen *read*, nemlig elevenes prestasjoner.




Figur 4.5: Scatterplot av teacher_exp og teacher_age

Fra figur 4.5 avdekkes det åpenbare falske datapunkter. Den kategoriske gruppen med verdi *1* for variabelen tilknyttet lærerens alder, *teacher_age*, representerer lærere under 25 år. Det vil ikke være mulig å ha over 20 års erfaring som lærer og samtidig være yngre enn 25 år. Grunnet et relativt stort datasett vil det dog ikke ha noen synlig effekt for modellen at disse observasjonene fjernes.

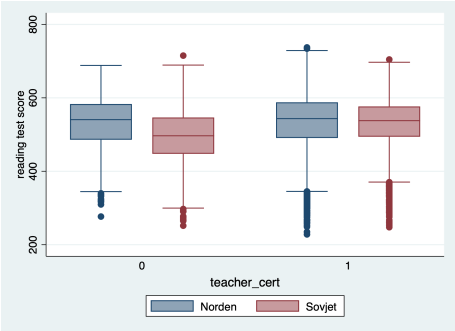
### Kategorisering av variabler

En numerisk kodet variabel som brukes for å markere kategorier kalles for en *dummyvariabel*. Dummyvariabler benyttes for å lage kategorivariabler som har klare økonomiske tolkninger som kan brukes direkte i de økonometriske modellene. I regresjonsanalysen er det irrelevant om en person er fra Latvia, Russland eller Moldova, så det genereres en ny binær variabel ved navn *sovjet*. Alle land som tidligere har vært en del av Sovjetunionen får verdien 1, mens landene i Norden får verdien 0, som vist i tabell 4.6. Denne variabelen opprettes for å kunne analysere hvilke faktorer tilknyttet læreren som påvirker leseferdighetene til en elev i Norden sammenlignet med en elev i Øst-Europa.

| Variabel | Tilstand | Verdi | Dummyvariabel |
|---|---|---|---|
| idcntry | 428 | Latvia | sovjet lik 1 |
| | 498 | Moldova | |
| | 643 | Russland | |
| | 352 | Island | sovjet lik 0 |
| | 578 | Norge | |
| | 752 | Sverige | |

Tabell 4.6: Dummyvariabel for variabelen sovjet.

Med den nye variabelen er det enklere å se på interessante sammenlikninger mellom Norden og det tidligere Sovjetunionen. Tidligere ble fordelingen av leseferdigheter med hensyn på lærerens sertifisering presentert i figur 4.1. Med den nye variabelen vil forskjellene for de to grupperingene komme tydeligere frem, som vist i figur 4.6. Differansen mellom mediankarakteren representert ved streken i sentrum av boksen for en lærer som er sertifisert, versus en ikke-sertifisert lærer, er betydelig større for Sovjet enn Norden. Samtidig viser figuren at sovjetiske elever med sertifisert lærer har en mindre spredning av resultater.




Figur 4.6: Fordelingen av elevers leseferdigheter ut i fra lærerens sertifisering, separert for Norden og Sovjet.

### 4.2.4 Valg av kontrollvariabler

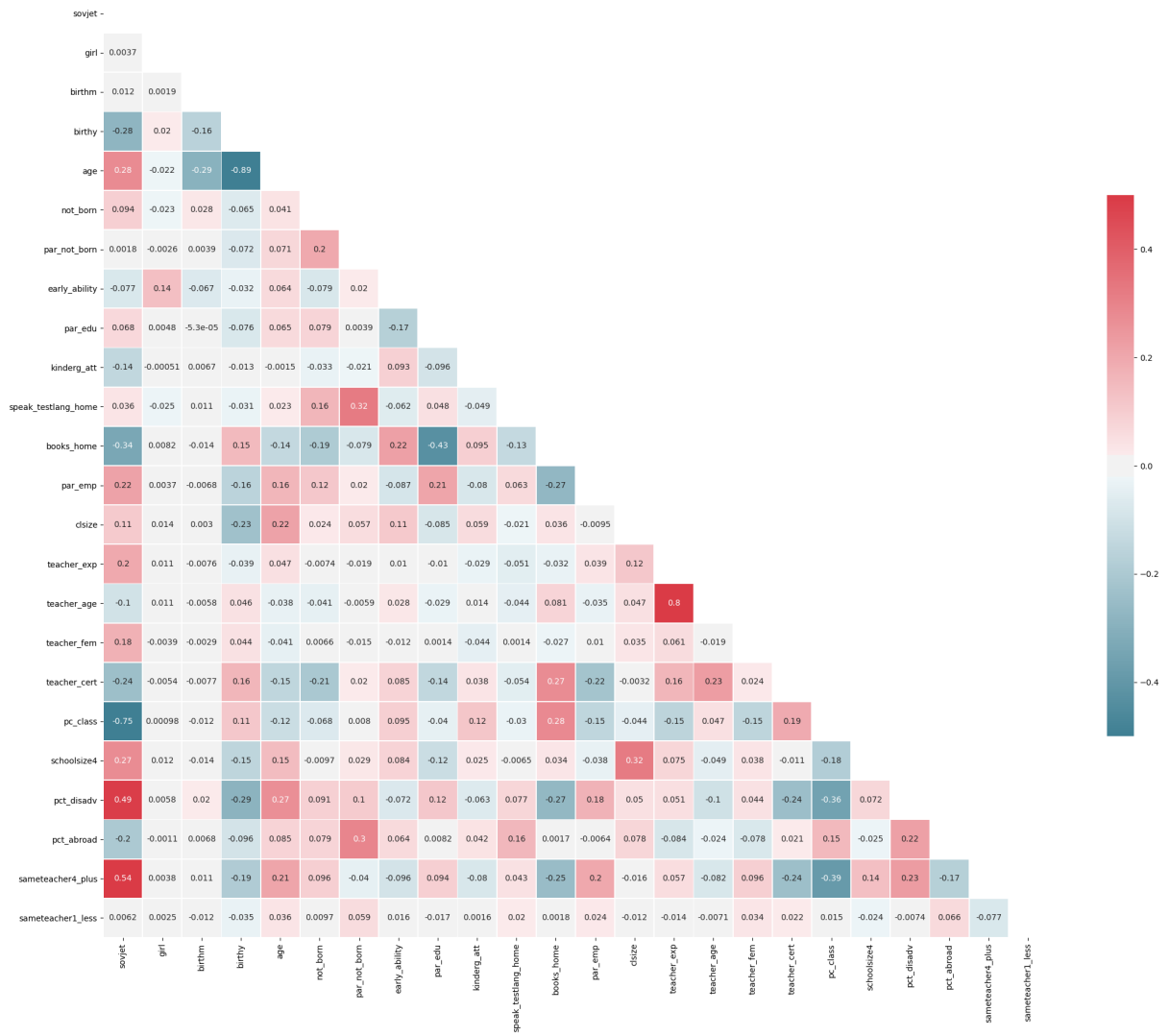
Kontrollvariabler legges til for å unngå utelatt-variabel problemet. Ved å inkludere flere forklaringsvariabler enn kun de tilknyttet læreregenskaper i datasettet, får man muligheten til å se om effekten av hver enkelt lærer-relaterte variabel beholdes når det inkluderes flere forklaringsvariabler. Slik kontrolleres det for effekten av de andre forklaringsvariablene. Et viktig utgangspunkt når man velger disse kontrollvariablene er å ha representanter for hver av kategoriene i skoleproduksjonsfunksjonen. Faktorer tilknyttet familie- og elevkarakteristika, medelevkarakteristika og skolefaktorer bør derfor være representert blant kontrollvariablene.

Det er fortsatt behov for å redusere størrelsen på datasettet før det brukes til videre analyser, da for mange variabler vil kunne medføre multikollinearitet og gi en modell som er vanskelig å tolke. Deskriptiv statistikk for de mulige kontrollvariablene er vist i figur 7.3 i appendix. For å få en så presis modell som mulig er det viktig å ha stort nok datagrunnlag av variablene som skal brukes i regresjonen. Det mangler datapunkter for foreldrenes inntekt i Russland og Latvia, noe som medfører at denne variabelen utelukkes. Skolens lokasjon utelukkes også grunnet manglende datagrunnlag, da omtrent 25% av observasjonene mangler denne verdien.

Variabelen *idgrade* inneholder informasjon om hvilken klasse elevene går i. Datasettet er opprinnelig kun for elever fra fjerde-klasse, men det inneholder også noen elever fra tredje-klasse. Denne variabelen utelukkes, da den ikke er relevant for undersøkelsene i denne oppgaven. Dette kan også anses som misvisende informasjon, da datasettet opprinnelig er ment for fjerde-klassinger. Det er i midlertidig kun Russland som har data for tredje-klassinger. Variabelen anses ikke som relevant for analysen og tas derfor ikke med som en kontrollvariabel.

#### Korrelasjonsmatrise og multikollinearitet

Hvis noen av variablene korrelerer sterkt med hverandre er de å anse som overflødige, ved mindre de er av spesiell interesse av andre årsaker. Skulle det skje at to variable har en perfekt korrelasjon, vil vi ha et tilfelle av multikollinearitet. Dette bryter med forutsetningene for multippel lineær regresjon om at de uavhengige variablene skal være uavhengige. En oversikt over korrelasjonen mellom variable er vist i figur 4.7. Hver rute i figuren viser korrelasjonskoeffisienten mellom to variable, som forklart i kapittel 3.3.




Figur 4.7: Korrelasjonsmatrise for foreløpig utvalgte kontroll- og interessevariabler

Utfra korrelasjonsmatrisen kan man se at året en elev er født og alderen, representert ved henholdsvis variablene *birthyear* og *age*, er svært korreleterte. Dette bekreftes utfra VIF-scoren i figur 7.4. Dette gir mening, da fødselsåret naturlig nok påvirker alderen til en person. Alderen til en elev beholdes som kontrollvariabel da det er gitt som en kontinuerlig variabel, mens fødselsåret forkastes. Antall bøker en elev har i hjemmet og foreldrenes utdanning viser en viss grad av korrelasjon, i tillegg til at de representerer noen av de samme komponentene i skoleproduksjonsfunksjonen. Disse egenskapene er gitt ved henholdsvis variablene *books_home* og *par_edu*. Kun foreldrenes utdanning tas med videre i analysen, da det antas at denne variabelen har høy forklaringskraft, samt at det er interessant å se om foreldre med høyere utdanning påvirker fokuset av lesing i hjemmet. Denne kategoriske variabelen for foreldres utdanning gjøres om til en binær variabel for høyere utdanning, slik at man får tydeligere frem om foreldrene har fullført det man i Norge anser for å være høyere utdanning. Prosesseringen av denne variablene er vist i tabell 4.7.

| Variabel | Tilstand | Verdi | Dummyvariabel |
|---|---|---|---|
| teacher_edu | University degree | 1 | higher_edu lik 1 |
| | Post secondary (not university) | 2 | |
| | Upper secondary | 3 | higher_edu lik 0 |
| | Lower secondary | 4 | |
| | Not completed lower secondary | 5 | |

Tabell 4.7: Dummyvariabel for variabelen *teacher_edu*. Det lages en variabel med navn *higher_edu* med verdien 1 hvis foreldrene har fullført høyere utdanning.

Den prosentvise andelen av antall elever fra økonomisk vanskeligstilte hjem på skolen, gitt ved variabelen *pct_disadv*, korrelerer med variabelen for de Østeuropeiske landene med en korrelasjonsfaktor på 0.49. En tommelfingerregelen er at to variabler har en sterk sammenheng dersom korrelasjonskoeffisienten er over 0.8 og en moderat sammenheng dersom den er ca. 0.5. Korrelasjonsfaktoren på 0.49 utviser derfor høy nok grad av korrelasjon til at variabelen forkastes.

Det er som sagt viktig å ta med kontrollvariabler fra de ulike kategoriene i skoleproduksjonsfunksjonen. For elevens karakteristikk, så anses både kjønnet, alderen, foreldrenes bakgrunn, om man snakker testspråket i hjemmet og tidligere leseferdigheter som viktige. Kjønn er interessant for å kunne se hvordan kjønnet på læreren kan spille inn på leseferdigheter til barna avhengig av deres kjønn. Måneden en elev er født hadde en VIF-score på 64.59, noe som viser en høy grad av korrelasjon med andre variable. Ettersom alle elever som er født i samme år begynner i samme trinn, i tillegg til den høye VIF score, så utelukkes måneden eleven er født (*birthm*), da de fleste i samme alder har gått like lenge på skolen. Om eleven er født i landet og om man snakker språket hjemme er relatert til språkbarrieren til studenten. Om eleven snakker testspråket i hjemmet, gitt ved variabelen *speak_testlang_home*, velges ut som kontrollvariabel, da det anses som utslagsgivende at eleven snakker testspråket til vanlig.

Det anses også som viktig om studenten hadde gode leseferdigheter som barn. Variabelen *early_ability* beholdes derfor som kontrollvariabel, mens om eleven var i barnehagen eller ikke (*kindberg_attend*) anses ikke som utslagsgivende. Som vist i korrelasjonsmatrisen, 4.7, så avhenger både antall bøker i hjemmet, inntekt, foreldrenes arbeidsstatus og foreldrenes utdannelse av hverandre. Alle disse variablene er tilknyttet foreldrene til elevene. Kun foreldrenes utdannelse beholdes som kontrollvariabel, da de andre viser grad av linearitet, da de trolig har en kausalitet.

For skolens karakteristikk beholdes alle variablene: klassestørrelse, om det er en PC tilgjengelig i klasserommet og antall studenter født i et annet land som er elever på skolen. Oppsummert så er det tatt høyde for korrelasjon, manglende data, skoleproduksjonsfunksjonen, relevante sammenhenger mellom lærer og læring (2.2 funnet i tidligere studier, samt skjønn for å komponere variabelsammensetningen som vist i tabell 4.8.

### Korrelasjon blant interessevariabler

Lærerens alder og erfaring, gitt ved henholdsvis variablene *teacher_age* og *teacher_exp*, viser svært høy korrelasjon. Det er ikke overraskende at korrelasjonen er høy, da antall års erfaring som regel henger nøye sammen med alderen til en person. Begge variablene

| **Variabel** | **Forklaring** | **Forkortelse** |
|---|---|---|
| Tidlig evne til å lese | Lik 1 hvis sann. | early_ability |
| Alder | Kontinuerlig variabel | age |
| Foreldrenes utdanning | 5 ulike kategorier | par_edu |
| Kjønn | Lik 1 hvis jente | girl |
| Snakker testspråket i hjemmet | Lik 1 hvis sann | speak_testlang_home |
| Klassestørrelse | Kontinerlig variabel | clsize |
| PC i klasserommet | Lik 1 hvis det er en PC i klasserommet | pc_class |
| Prosent av antall studenter født i et annet land på skolen | 4 ulike kategorier | pct_abroad |

Tabell 4.8: Forklaring av kontrollvariabler sammen med en forkortelse for hver variabel som er brukt i regresjonsanalyser. Kontrollvariablene er variablene referert til som $X_i$ i tabell 4.1.

beholdes for videre analyser, da måten de korrelerer kan være forskjellig for Norden og Sovjet. Dette er noe som kan være utslagsgivende i regresjonsmodellen, da mange av de eks-sovjetiske landene har opplevd en resesjon etter oppløsningen av Sovjetunionen som har ført med seg økonomiske nedgangstider og økt arbeidsledighet. Dermed kan det ikke være usannsynlig å finne lærere med høy alder og lite arbeidserfaring som lærer i Sovjet. Det er også verdt å merke seg at *Sameteacher4_plus* og *sovjet* korrelerer, men beholdes for å anvendes i lærer-analysen.

## 4.3 Fordeler og begrensinger ved datasettet

Måling og standarder for datagrunnlaget kan variere mellom de ulike landene. Når hovedformålet med oppgaven er å undersøke avvik mellom ulike lands elevpresentasjoner, er det nødvendig med tydelige standarder for alle variablene som skal kunne tolkes. Hvis denne standarden ikke er tydelig, vil det forekomme avvik i den innsamlede dataen. Med tanke på at informasjonen om lærere er hentet inn i form av spørreundersøkelser vil dette innebære en usikkerhet til de endelige resultatenes gyldighet. Det har blitt funnet funnet feilrapporteringer som ikke har blitt fjernet fra datagrunnlaget, slik den som fremkommer i figur 4.5.

Vedrørende elevprestasjoner så er det mulig at lesetestens resultater kan bli påvirket av om nasjonene er flerspråklig, med tanke på at det er elevenes evnenivå man håper at testen skal representere.

En annen usikkerhetsfaktor som må tas høyde for er at man ikke vet hvilken utdanning lærerne har, derav manglende informasjon om *teacher_edu*-variabelen. I dette tilfellet er lærernes faglige og pedagogiske kvalifikasjoner og evner ukjent. Om en ytterligere optimering av ressurser kan påvirke elevpresentasjoner forblir da ukjent. Datasettet tar derimot for seg lærerens erfaring og sertifisering.

For å gi et helhetlig bilde av hvilke av faktorer tilknyttet læreren som påvirker elevenes leseferdigheter, så skulle et ideelt datasett også blant annet inkludert felt for modenhet, hvordan tilpasset undervisning utføres, antall elever per lærer, om elevene følger bestemte studielinjer gitt individuell motivasjon og interessefelt, opptakskrav, ekstern finansiering, arbeidsetterspørsel og egeninnsats.

En av fordelene ved datasettene er at de inneholder et stort utvalg av observasjoner, noe som kan jevne ut eventuelle ekstremobservasjoner mot et gjennomsnittlig og normaldistribuert nivå. En annen fordel er at datasettet fanger opp hvor mange år en elev har vært undervist av den samme lærer, noe som ifølge Hanuschek (2020) er av stor interesse, da en elev som har hatt samme lærer over flere år viser høyere prestasjoner enn en elev som ikke har hatt en kontinuerlig undervisning av samme lærer, som nevnt i kapittel 2.2.

Datasettet viser et klart brudd på en av forutsetningene forklart i kapittel 3.2. Variansen til støyleddet er antatt konstant for at standardfeil ikke skal oppstå, noe som kalles for homoskedastisitet. Dersom restleddet ikke har konstant varians, så omtales det som heteroskedastisitet: $Var(\epsilon_i) \neq \sigma^2$. Brudd på denne antagelsen kan føre til forvrengning av funn og dermed svekke analysen. Heteroskedastisitet ble undersøkt ved hypotestester i Stata, hvor det ble formulert nullhypotester hvor det antas homoskedastisitet og alternativhypoteser for heteroskedastisitet. Figur 4.8 viser resultatene for en test gjennomført for en regresjonsmodell med alle variablene i datasettet i Stata. P-verdien er 0 for alle variable og kjikvadrat-verdien er høy, noe som medfører at nullhypotesen forkastes. Dette betyr at noen av forklaringsvariablene påvirker variansen, men dette tas ikke høyde for i videre analyser, da effekten av det anses som neglisjerbar.

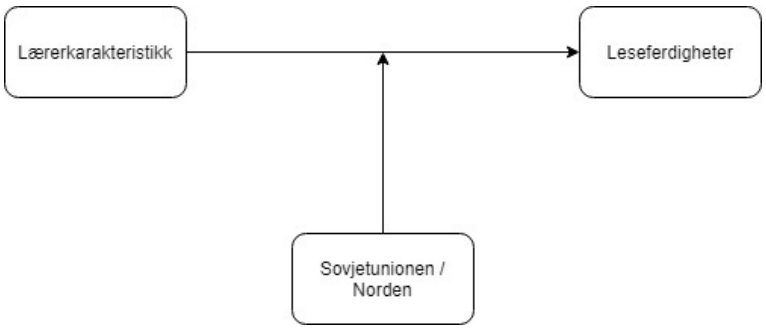
```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of read

         chi2(1)      =   235.46
         Prob > chi2  =   0.0000
```

Figur 4.8: Test av heteroskedastisitet.

**Kapittel 5**

# Regresjonsanalyse

For å undersøke hvilke faktorer tilknyttet læreren som påvirker leseferdighetene til en elev i Norden sammenlignet med en elev i Øst-Europa, så er det først interessant å undersøke hvilke faktorer tilknyttet læreren som påvirker leseferdighetene til en elev for et datasett som inkluderer begge to de regionene. Analysene er derfor delt inn i to deler, som illustrert i figur 5.1. Del én undersøker om lærerkarakteristikken har en effekt på leseferdighetene, mens del to undersøker om denne relasjonen påvirkes av at en elev bor i Norden eller Øst-Europa.




Figur 5.1: Det undersøkes nærmere om lærerkarakteristikken har forskjellig påvirkning på en elev i Norden sammenlignet med en land som tidligere har vært medlem av Sovjetuionen.

## 5.1 Lærerens påvirkning på leseferdighetene

I første del undersøkes det om læreren kan sies å ha en antatt effekt på elevens leseferdigheter, eller om dette er en påstand som kan falsifiseres.

### 5.1.1 Regresjonsanalyse og hypotesetesting

For å se på lærerens påvirkning på leseferdighetene benyttes minste kvadraters metode som beskrevet i kapittel 3.1. Ved å benytte denne metoden lages det en modell som viser sammenhengen mellom leseferdigheten til en elev og de relevante variablene. For å se om hver enkelt lærer-variabel har en påvirkning på leseferdighetene, så sjekkes hver enkelt lærervariabel opp mot leseferdighetene. Dette gjøres ved å estimere en simpel regresjonsmodell for hver eneste lærer-variabel, hvor den enkelte lærer-variabelen er den eneste avhengige variabelen, som vist i ligning (5.1).

$$\begin{aligned} read &= \beta_0 + \beta_1 teacher\_fem + \epsilon \\ read &= \beta_0 + \beta_1 teacher\_exp + \epsilon \\ read &= \beta_0 + \beta_1 teacher\_age + \epsilon \\ read &= \beta_0 + \beta_1 teacher\_certificate + \epsilon \\ read &= \beta_0 + \beta_1 sameteacher\_1less + \epsilon \\ read &= \beta_0 + \beta_1 sameteacher\_4plus + \epsilon \end{aligned} \tag{5.1}$$

Deretter sjekkes det om hver variabel er signifikant. Dette sjekkes ved en enkel hypotesetest som beskrevet i 3.4 for hver variabel. Det formuleres en nullhypotesten der koeffisienten foran hver interessevariabel er lik 0, det vil si om leseferdighetene er uavhengig av interessevariabelen. Nullhypotesen forkastes dersom p-verdien er under det bestemte signifikansnivået som er satt, hvor det i denne rapporten benyttes et nivå på 0.05. Nullhypotesen forkastes dersom sannsynligheten er tilstrekkelig lav til at koeffisienten kan sies å være lik 0. Om man forkaster nullhypotesen sies variabelen å være statistisk signifikant. Resultatet fra en hypotestetest for om en kvinnelig lærer har påvirkning på leseferdighetene er vist i figur 5.2. P-verdien er **0.2086**, hvor variabelen dermed ikke beregnes som signifikant for modellen med et signifikansnivå på 5%.

```
. test teacher_fem

 ( 1)  teacher_fem = 0

       F(  1, 24319) =    1.58
            Prob > F =    0.2086
```

Figur 5.2: Hypotesetesting

For å automatisere prosessen med nullhypoteser, så brukes *outreg2*-kommandoen i Stata. P-verdien kan dermed leses ut direkte sammen med regresjonskoeffisientene. Resultatene fra regresjonen er vist i Tabell 5.1. Resultatene viser koeffisientene foran hver variabel i regresjonsmodellen, mens parentesen under koeffisienten forteller p-verdien til hver koeffisient.

Deretter formuleres en lineær grunnmodell med kontrollvariabler for å se om kontroll variablene kan ha påvirkning på leseferdighetene. Grunnen til at vi gjør dette er for å sørge for at det faktisk er en relasjon mellom lærervariablene og leseferdighetene. Ved å legge til kontrollvariabler i modellen som forventes å ha en effekt på den avhengige variabelen, så testes det om effekten av variabelen endrer seg når kontrollvariablene er inkludert i modellen, slik at man kan utelukke at sammenhengen ikke skyldes tredje-variabler utelatt fra analysen. Regresjonsmodellene blir nå som vist i ligning (5.2), hvorav $X_i$ inkluderer alle kontrollvariablene som er valgt i kapittel 4.2.4. Resultatene fra disse analysene er vist i tabell 7.1 i appendix, hvor det er kjørt en analyse for hver interessevariabel. Tilsvarende skrives p-verdiene for hver av koeffisientene direkte til resultatene med regresjonskoeffisientene. P-verdiene finnes ved å gjøre tilsvarende hypotesetest som er gjort i figur 5.2.

$$read = \beta_0 + \beta_1 teacher\_fem + X_i + \epsilon \tag{5.2}$$

### 5.1.2 Empiriske resultater

Basert på resultatene i Tabell 5.1 for regresjonsmodeller uten kontrollvariabler, så er alle variabler signifikante utenom en. Basert på hypotesetesten for lærerens kjønn, *teacher_female*, så beholdes nullhypotesten og variabelen regnes derfor ikke som signifikant. Basert på resultatene fra regresjonsanalysen med kontrollvariabler i tabell 7.1, så bekreftes det at kjønnet til en lærer ikke kan bevises å ha en påvirkning på leseferdighetene til en elev, da p-verdien er på 0.181. Lærerens kjønn benyttes dog i videre analyser, da det er interessant

| VARIABLES | (1) read | (2) read | (3) read | (4) read | (5) read | (6) read |
|---|---|---|---|---|---|---|
| 2.teacher_age | 35.63 | | | | | |
| | (0) | | | | | |
| 3.teacher_age | 32.82 | | | | | |
| | (0) | | | | | |
| 4.teacher_age | 29.05 | | | | | |
| | (0) | | | | | |
| 5.teacher_age | 39.53 | | | | | |
| | (0) | | | | | |
| 6.teacher_age | 40.73 | | | | | |
| | (0) | | | | | |
| teacher_exp | | 0.232 | | | | |
| | | (2.50e-08) | | | | |
| teacher_fem | | | -1.947 | | | |
| | | | (0.209) | | | |
| teacher_cert | | | | 30.26 | | |
| | | | | (0) | | |
| sameteacher1_less | | | | | 16.73 | |
| | | | | | (0.000303) | |
| sameteacher4_plus | | | | | | -20.88 |
| | | | | | | (0) |
| Constant | 497.2 | 526.0 | 531.8 | 503.8 | 530.0 | 537.8 |
| | (0) | (0) | (0) | (0) | (0) | (0) |
| Observations | 24,307 | 23,679 | 24,321 | 24,136 | 24,069 | 24,069 |
| R-squared | 0.011 | 0.001 | 0.000 | 0.021 | 0.001 | 0.019 |

pval in parentheses

Tabell 5.1: Regresjonsanalyse med alle interessevariabler uten kontrollvariabler.

om en kvinnelig lærer har påvirkning på en elev i Sovjetunionen eller Norden, selv om det ikke har påvirkning i en generell modell.

Lærerens alder viser seg signifikant både med og uten kontrollvariabler. Kategorien for lærere under 25 år brukes som referansekategori. Alle de andre kategoriene har en koeffisient mellom 19 og 26, noe som vil si at en eldre lærer tyder på å utgjøre en effekt på lesescoren. Lesescoren øker ikke lineært med lærerens alder, men det observeres at det er lærerne i den høyeste aldersgruppen som har den høyeste koeffisienten, noe som vil si at denne kategorien har den største effekten på leseferdighetene. Hvis en elev har en lærer fra den eldste lærergruppen, så vil dette være med å bidra til en økt poengsum på 26.33 poeng sammenlignet med hvis læreren hadde vært i kategorien under 25 år.

Lærerens erfaring er signifikant med en p-verdi tilnærmet lik null i modellen uten kontrollvariabler, mens verdien øker til 0.015 i modellen med kontrollvariabler. Variabelen er fortsatt signifikant, med en positiv koeffisient på ca. 0.1. Dette vil si at en lærer med 30 års erfaring vil bidra med å øke lesescoren til en elev med ca. 3 poeng. Dette utgjør ikke en stor forskjell, men det er interessant å bemerke seg at en lærer sin erfaring ikke

har et stort bidrag til en elev sine leseferdigheter. Det er uansett overraskende hvor liten positiv effekt erfaring har. Om lærerens alder og erfaring begge er tegn på modenheten, refleksjon og kunnskap, så ville det forventes at de begge hadde omtrentlig samme effekt. Man kunne også forventet at alder ville ha en mindre positiv effekt enn erfaring, som følge av de aspektene ved alderdom som påvirker jobbhverdagen, slik som energinivå og økende grad av helseproblemer.

Videre kan det slås fast at lærerens sertifisering er signifikant og nullhypotesen må forkastes. Om læreren er sertifisert så er dette estimert å utgjøre en 16.7 poeng forskjell i elevens leseferdigheter. Vedrørende skolepolitikken, så er det interessant med resultatene for en elev som har hatt samme lærer i under ett år og mer enn fire år. Begge variablene er signifikante i begge modellene, hvorav koeffisientene er gitt ved 11.69 og -12.76 for henholdsvis å ha samme lærer i mindre enn et år og mer enn 4 år. Det å ha samme lærer i maksimalt ett år har en positiv effekt med nesten 12 poeng på leseferdigheter, mens det å ha samme lærer i minimum fire år vil ha en negativ effekt med omtrent samme verdi.

Man skulle tro at det å ha samme lærer over lengre tid bidrar til å øke leseferdighetene til en elev, da læreren lærer seg å kjenne eleven bedre, men dette viser seg å være noe som ikke nødvendigvis stemmer. Dette står i strid med tidligere studier, ref. kapittel 2.2.

## 5.2 Lærerens påvirkning på leseferdighetene i Sovjetunionen sammenlignet med Norden

Utfra resultatet i forrige underkapittel er det nå relevant å undersøke om lærere påvirker elevenes leseferdigheter på forskjellige måter i Norden sammenlignet med land som tidligere var en del av Sovjetunionen. Det skal nå undersøkes om det kan konstanteres en reell forskjell mellom regionene eller om dette kan falsifiseres.

### 5.2.1 Lineær regresjon og hypotestetesting

For å se om Sovjetunionen har en påvirkning på leseferdighetene til elevene, så kjøres en enkel regresjonsmodell, med og uten kontrollvariabler, der variabelen *sovjet* er den eneste interessevariabelen. Regresjonsmodellene er vist i ligning (5.3), mens regresjonskoeffisientene og p-verdiene til hver koeffisient for de to ulike regresjonsmodellene er vist i tabell 5.2.

$$\begin{aligned} read &= \beta_0 + \beta_1 sovjet + \epsilon \\ read &= \beta_0 + \beta_1 sovjet + X_i + \epsilon \end{aligned} \tag{5.3}$$

Deretter undersøkes effekten av hva slags innvirkning lærerkarakteristikken har for en elev i Norden sammenlignet med Sovjetunionen. Dette gjøres ved å legge til et interaksjonsledd mellom *sovjet*-variabelen og de ulike interessevariablene. Regresjonsmodellen blir dermed ikke-lineær grunnet interaksjonsleddet. Dette gjør det mulig å se f.eks. hva slags påvirkning kjønnet til læreren har å si for leseferdighetene til en elev i Sovjetunionen sammenlignet med Norden. Det lages en ikke-lineær modell med interaksjonsledd mellom *sovjet*-variabelen og hver av interessevariablene, og det kjøres en analyse med og uten kontrollvariabler som vist

| VARIABLES | (1)<br>read | (2)<br>read |
|---|---|---|
| sovjet | -12.15 | -10.54 |
| | (0) | (0) |
| early_ability | | 22.05 |
| | | (0) |
| age | | 17.26 |
| | | (0) |
| higher_edu | | 33.31 |
| | | (0) |
| girl | | 13.40 |
| | | (0) |
| speak_testlang_home | | -21.29 |
| | | (0) |
| clsize | | 0.630 |
| | | (0) |
| pc_class | | 7.696 |
| | | (5.86e-08) |
| pct_abroad | | -2.886 |
| | | (0.00213) |
| Constant | 535.4 | 282.5 |
| | (0) | (0) |
| Observations | 24,979 | 18,842 |
| R-squared | 0.007 | 0.255 |

p-value in parentheses

Tabell 5.2: Resultater for regresjonsmodeller med og uten kontrollvariabler for *sovjet* som variabel.

i ligning (5.4) og (5.5). Hypotesetester kjøres automatisk for hver koeffisient som tidligere, og resultatene er vist i tabell 7.2, 7.3, 7.4 og 7.5.

$$read = \beta_0 + \beta_1 sovjet + \beta_2 teacher\_fem + \beta_3 sovjet \cdot teacher\_fem + \epsilon \tag{5.4}$$

$$read = \beta_0 + \beta_1 sovjet + \beta_2 teacher\_fem + \beta_3 sovjet \cdot teacher\_fem + X_i + \epsilon \tag{5.5}$$

### 5.2.2 Empiriske resultater

For å validere at det faktisk eksisterer en strukturell endring mellom Sovjetunionen og Norden kan man kjøre Chow-testen. For å finne en strukturell ulikhet mellom Nordens og Sovjetunionens lærere, der Norden er referansegruppen, testes to modeller mot hverandre. Den ene for lærere i Norden, og den andre for lærere i Sovjet. Nullhypotesen formuleres til at det ikke eksisterer en strukturell endring i dataen avhengig av om det er en lærer fra Norden versus Sovjet, hvilket Chow-testen falsifiserer, som fremkommer av testen i avsnitt

7.3 i appendix. En annen måte å teste det på er ved å se på hver enkelt koeffisient og teste den enkeltvis.

Å være et tidligere sovjetisk land har innvirkning på leseferdighetene, basert på resultatene i tabell 5.2. Nullhypotesen om at koeffisienten til variabelen *sovjet* ikke har påvirkning på leseferdighetene forkastes, da p-verdien er 0.00 både for regresjonsmodellen med og uten kontrollvariabler. Koeffisientene foran hver forklaringsvariabel uttrykker effekten av forklaringsvariabelen på leseferdighetene. Regresjonsmodellen med kontrollvariabler er en langt bedre modell enn uten, da vi kan se $R^2$-scoren gikk fra 0.007 til 0.255. Denne verdien er dog ikke helt sammenlignbar, da man må benytte Adjusted $R^2$ verdien for å ta hensyn til at flere variabler inkluderes i modellen. Koeffisienten til sovjet tilsier at en elev fra Sovjetunionen har en lesescore som er 10.54 lavere enn en elev i Norden. Gjennomsnittsverdien for leseferdigheter i Sovjetunionen er 533.90, mens gjennomsnittsverdien er 535.36 i Norden. Differansen er dermed noe høyere i den estimerte modellen enn det er i virkeligheten, noe som gjenspeiles i at $R^2$-scoren er langt unna den perfekte verdien, nemlig 1.

Basert på t-tester presentert i tabell 7.2 for modell 2 fremkommer det at det kun er lærere i alderen '40-49' og '60 eller mer' at det er en signifikant forskjell for Sovjet vs. Norden. For lærere i kategorien 40-49 år vil effekten påført lesescore av å være fra sovjet være summen av interaksjonsvariabelen og forklaringsvariablen for lærers alder som kombineres i interaksjonsvariabelen med *sovjet*. Effekten av at den sovjetiske læreren er i den gitte aldersgruppen utgjør dermed 27.8 poeng, til sammenlikning med en lærer i alderen 'minst 25 år'. Dette er en differanse på 12.8 poeng fra lærere fra samme aldersgruppe i Norden, som kun vil utgjøre en 15 poengs forskjell på sine elever. For en lærere fra Sovjet utgjør det så mye som 42.3 poeng effekt av å tilhøre den eldste aldersgruppen, til forskjell fra den yngste aldersgruppen. Differansen mellom Sovjet og Norden utgjør 26 poeng, ergo er det enda mer gunstig å ha en gammel lærer i Sovjet til sammenlikning med Norden.

Fra modell 2 i tabell 7.3 fremkommer det at lærerens kjønn og erfaring har en statistisk signifikant betydning for å utgjøre en forskjell mellom Norden og Sovjet, mens man ikke har statistisk grunnlag for å si at erfaring alene er signifikant i denne modellen. Dermed kan erfaring antas å ha en mindre betydning for en lærer fra Norden enn fra Sovjet. Effekten erfaringen til læreren gir for en lærer i Sovjet har en marginal effekt på 0.35. Tilsvarende for en lærer med 40 års erfaring vil dermed lesescoren øke med 14 poeng, til forskjell fra en lærer uten erfaring, eller en lærer fra Norden, uavhengig av erfaring.

Fra modell 4 i tabell 7.3 fremkommer det at lærerens kjønn er utslagsgivende når man separerer for de to gruppene med land, til tross for at modellen tilknyttet den generelle påvirkningen lærere har på sine elever viste at lærerens kjønn var estimert til å være ubetydelig for Norden og Sovjet samlet. Om man er en kvinnelig lærer fra Sovjet er det forventet å øke leseferdighetene med 12.8 poeng. Til motsetning vil ikke lærerens kjønn utgjøre en forskjell i Norden.

Sertifisering er også signifikant forskjellig for Norden og Sovjet, som vist i modell 2 tabell 7.4, med en marginaleffekt lik 26.7 for en sertifisert sovjetisk lærer, til sammenlikning med en usertifisert lærer. Som for kjønn, så vil vil ikke sertifisering heller utgjøre en forskjell i Norden.

Fra modell 2 i tabell 7.5, så fremkommer det at for elever som har læreren i opptil ett år er signifikant forskjellig for Norden og Sovjet med en betydelig differanse. I Sovjet vil

dette utgjøre -6 poeng på lesescoren, mens i Norden vil det øke lesescoren med 29.3 poeng. Det er altså en invers relasjon mellom hvordan leseferdigheter påvirkes av at man har hatt en lærer i opptil ett år for Norden og Sovjet. Effekten av at læreren er sovjetisk utgjør dermed en forskjell på -35.3 poeng, og utgjør en stor forskjell mellom de to landegruppene funnet i denne oppgaven.

Fra modell 4 i samme tabellen fremgår det at det utgjør en signifikant forskjell å ha en lærer i minst fire år i Norden versus Sovjet. For en Sovjetisk elev vil effekten av å ha en lærer i minst fire år utgjøre -4.8 poeng, mens det vil utgjøre -20.4 poeng for en Nordisk elev. Altså er effekten 4 ganger så stor for en nordisk elev, da differansen er på 15.6 poeng.

Generelt er ikke koeffisientene sammenlignbare mellom modellene, men de gir en indikasjon på om effektene er positive, negative og store eller små for leseferdighetene til en elev. Oppsummert kan man si at blant lærere i sovjet teller det positivt å være kvinne, av høy alder, ha mye erfaring og en sertifisering. Hvor lenge man er lærer for en elev er ikke en personlig egenskap, og kan brukes sidestilt funnene om lærerens karakteristika.

For Norden vil kun alder ha en bevist betydning i følge våre modeller. Utifra disse modellene er det dermed vanskeligere å lage policymodeller for det nordiske utdanningssystemet, enn hva det vil være for det sovjetiske utdanningssystemet, basert på våre resultater.

### 5.2.3 Sammensetningen av læreregenskaper for lærere i Norden vs. Sovjet

Ettersom det er avdekket forskjeller mellom lærerne avhengig av hvilken gruppering av land de kommer fra, er det interessant å se på sammensetninger av egenskaper som lærerne har. Spesielt er det interessant å se hvordan de variablene som har vist seg å være signifikante og svært betydningsfulle sammenfaller hos lærerne. Eksempelvis: Er det mer sannsynlig at en lærer er fra Sovjet om den er gammel, sertifisert og kvinne? Eller vil det være høyere sannsynlighet for å være fra Norden om man er under 25, har noen år erfaring, og elever som presterer blant de 25% beste?

For å undersøke dette har datastrukturene blitt endret til at *read* er omgjort til en kategorisk variabel, der hver av de fire kategoriene utgjør 25% av elevene, rangert fra lavest til høyeste testresultat. Variablene som videre har blitt brukt har vært de som har blitt funnet signifikante gjennom regresjonsanalysen. Det har deretter blitt brukt mønstergjenkjenning ved hjelp av Apriori-algoritmen.. Lift er måleenheten som er brukt for å måle effekten til mønstrene, eller *reglene*, og forteller hvor ofte en regel inntreffer i en rad i datasettet sammen med en *konsekvens*, versus uten konsekvensen.

Et utvalg av resultater knyttet til om læreren er fra Sovjet eller Norden er blitt valgt ut for å underbygge argumentasjonen om at det trolig er store forskjeller mellom lærerne fra de to gruppene.

Reglene gir oss følgende informasjon fra den første raden: For en elev med lesescore blant de 25%, som ikke har hatt samme lærer i opptil ett år, som har en usertifisert lærer av eldste aldersgruppen vil dette være 2.94 ganger så sannsynlig at vedkommende som tok prøven er fra Sovjet. Det er altså langt flere usertifiserte lærere på over 60 år, som underviser elever de kun har hatt i opptil ett år med elever som scorer blant de 25% beste i Sovjet. Selve datagrunnlaget utgjør kun 0.34% av de 25000 observasjonene som er omtrent

| Rules | Rules head | Rules, support | Rules head, support | support | Confidence | Lift |
|---|---|---|---|---|---|---|
| read_4', 'sameteacher1_less_0.0', 'teacher_cert_0.0', 'teacher_age_6.0' | sovjet_1.0' | 0,0034 | 0,3406 | 0,0034 | 1.0 | 2,9361 |
| read_4', 'teacher_cert_0.0', 'teacher_age_6.0' | sovjet_1.0' | 0,0034 | 0,3406 | 0,0034 | 1.0 | 2,9361 |
| teacher_cert_0.0', 'teacher_age_6.0' | sovjet_1.0' | 0,0034 | 0,3406 | 0,0034 | 1.0 | 2,9361 |
| sameteacher1_less_0.0', 'teacher_cert_0.0', 'teacher_age_6.0' | sovjet_1.0' | 0,0034 | 0,3406 | 0,0034 | 1.0 | 2,9361 |
| sameteacher1_less_1.0', 'teacher_cert_1.0', 'teacher_age_2.0' | sovjet_0.0' | 0,0013 | 0,6594 | 0,0013 | 1.0 | 1,5165 |
| read_4', 'sameteacher1_less_1.0', 'teacher_age_2.0' | sovjet_0.0' | 0,0013 | 0,6594 | 0,0013 | 1.0 | 1,5165 |
| sameteacher1_less_1.0', 'teacher_age_2.0' | sovjet_0.0' | 0,0013 | 0,6594 | 0,0013 | 1.0 | 1,5165 |

Figur 5.3: Et utvalg av regler med høyest Lift-score for Sovjet og Norden

85 tilfeller, ergo er det ikke mulig å lage en generell regel, men det er nok til å kunne kalle det en tendens. 'Confidens' lik 1 betyr det at læreren er fra Sovjet ved absolutt alle tilfeller der læreren besitter alle disse egenskapene, samtidig som elevens leseferdigheter tilhører de 25 øverste prosentene.

En annen interessant sammenheng for Norden er funnet på rad seks: For en elev med lesescore blant de 25%, som har hatt samme lærer i opptil ett år, der læreren tilhører aldersgruppen 25-35 år, vil dette være 1.52 ganger så sannsynlig at vedkommende som tok prøven er fra Norden. Denne kombinasjonen av egenskaper er kun funnet i Norden, men er ikke beregnet å ha like stor sannsynlighet som den forrige reglen som ble presentert fordi den har en mindre datamengde til å støtte påstanden.

**Kapittel 6**

# Sammendrag og konklusjon

Ved å teste for forskjellige egenskaper tilknyttet lærere har det blitt funnet en klar sammenheng til hvordan elever presterer på leseundersøkelsen gjennomført på fjerdeklassinger fra Norge, Sverige, Island, Russland, Moldova og Latvia. Egenskapene som er studert er lærerens alder, erfaring, kjønn, sertifisering og om eleven har hatt den samme læreren i opptil ett år eller minst 4 år. Modellene som utgjør grunnlaget for evalueringen av dataen samlet inn er et resultat av multippel lineær regresjon, der leseferdigheter blir testet for avhengigheter til de nevnte læreregenskapene. Variabler tilknyttet de tre ulike kategoriene i skoleproduktfunksjonen er benyttet for å få en god og representativ modell med rom for økonometriske tolkninger.

Et generelt funn for alle landene er den viktige betydningen av lærerens alder og sertifisering. En sertifisert lærer og en høy alder har klare forbindelser med forbedring av elevens leseferdigheter. Lærerens påvirkningskraft på elevene er betydningsfull. Videre er det undersøkt hvordan nordiske lærere påvirker elevers leseferdigheter annerledes enn lærere fra tidligere sovjetiske stater. Forskjellen mellom Nordens kultur og økonomisk fremvekst er såpass stor mot land som fortsatt bærer spor av den kommunistiske superstaten som eksisterte frem til 1991, og etterlot seg økonomisk resesjon i landene som tar del i denne undersøkelsen. Det er avdekket store forskjeller for hvordan lærere fra de to grupperingene med land påvirker sine elever. For Sovjet har lærerens kjønn, alder, erfaring og sertifisering stor betydning, til forskjell fra Norden der kun lærerens alder kan bevises å utgjøre en forskjell.

Funnene vedrørende hvor lenge en elev har samme lærer tilsier at det er gunstig å bytte lærer hvert år i Norden, mens den administrative kostnaden tilknyttet dette mulig overgår effekten av utbyttet for utdanningssystemet i eks-Sovjetunionen. Uavhengig av land er det ikke å anbefale å ha samme lærer i minst fire år, da dette påvirker leseferdigheter negativt. Det er verdt å merke at dette ikke kan generaliserer for alle elever, men kun fjerdeklassinger, altså primært elever som har hatt samme lærer fra de begynte på skolen.

Hanushek konkluderte med at det å ha en 'god lærer' over flere år har en god effekt på elevens leseferdigheter, som nevnt i kapittel 2.2. De analysene gjort i oppgaven viser tendenser av det motsatte. Det er mulig at dette svaret er påvirket av at det er for få land tatt med i datagrunnlaget, men det kan også være en konsekvens av at man ikke vet hvem 'den gode læreren' er. Om datagrunnlaget i stor grad består av lærere som ikke utmerker seg som gode vil det dermed ikke bidra til å underbygge denne påstanden. Hvilke sammensetninger av gode egenskaper lærere har i Norden og Sovjet ble undersøkt, men gav ingen tydelige resultater som utpekte hverken Sovjet eller Norden for å ha flere gode lærere enn andre. Dette er noe som burde undersøkes videre for å kunne forstå hvilke ressurser som burde optimaliseres i satsning for å fremme lærerkvaliteten i de to landegruppene.

Det er vanskelig å optimalisere et utdanningssystem ut ifra hvilket kjønn som utgjør de beste lærerne, da man ønsker et mangfold i alle yrker man har, men en forståelse for hvorfor kvinnelige lærere i Sovjet tenderer å gi bedre resultater enn mannlige kan brukes til å undersøke hvilke egenskaper det er hos de forskjellige kjønnene som påvirker dette. Finnes det gode/dårlige vaner i gruppene som utgjør en forskjell for hvordan elevene lærer? Er mannlige lærere strengere enn kvinnelige? Er kvinnelige lærere generelt bedre på personlig oppfølging? Svar på disse spørsmålene kan avdekke forskjeller som kan korrigeres gjennom endret ledelse eller bevisstgjøring og dermed gi bedre læring for elevene.

**Kapittel 7**

# Appendix

## 7.1 Valg av variable

### 7.1.1 VSelect

Følgende kommando ble kjørt i Stata:

```
vselect read birthm age par_not_born par_edu kinderg_att books_home clsize
teacher_age teacher_cert pc_class school_loc~n pct_abroad sameteach~ss idgrade
girl birthy not_born early_abil~y income speak_test~e par_emp teacher_exp
teacher_fem teacher_edu schoolsize4 pct_disadv sameteach~us, best
```

Dette gir følgende output:

Optimal models:

| # Preds | R2ADJ | C | AIC | AICC | BIC |
|---|---|---|---|---|---|
| 1 | .0491734 | 3736.103 | 73079.57 | 73079.57 | 73093.11 |
| 2 | .1879624 | 2248.61 | 72060.8 | 72060.81 | 72081.12 |
| 3 | .2801688 | 1260.898 | 71282.82 | 71282.82 | 71309.91 |
| 4 | .2987711 | 1062.336 | 71114.6 | 71114.61 | 71148.47 |
| 5 | .3156673 | 882.1174 | 70957.96 | 70957.98 | 70998.61 |
| 6 | .3260333 | 771.9199 | 70860.31 | 70860.34 | 70907.73 |
| 7 | .335338 | 673.1261 | 70771.46 | 70771.49 | 70825.66 |
| 8 | .349226 | 525.2722 | 70635.99 | 70636.02 | 70696.95 |
| 9 | .3685811 | 318.9167 | 70441.85 | 70441.89 | 70509.59 |
| 10 | .3765401 | 234.6483 | 70360.87 | 70360.91 | 70435.38 |
| 11 | .3825977 | 170.7627 | 70298.76 | 70298.82 | 70380.05 |
| 12 | .3858909 | 136.4874 | 70265.19 | 70265.26 | 70353.25 |
| 13 | .3883194 | 111.4767 | 70240.58 | 70240.66 | 70335.42 |
| 14 | .3903318 | 90.92805 | 70220.28 | 70220.37 | 70321.89 |
| 15 | .3918395 | 75.7848 | 70205.28 | 70205.37 | 70313.66 |
| 16 | .3930855 | 63.44609 | 70193.02 | 70193.13 | 70308.18 |
| 17 | .3944141 | 50.22864 | 70179.85 | 70179.97 | 70301.78 |
| 18 | .3957737 | 36.68413 | 70166.32 | 70166.46 | **70295.03** |
| 19 | .3964812 | 30.11616 | 70159.75 | 70159.89 | 70295.23 |
| 20 | .3968824 | 26.82553 | 70156.45 | 70156.61 | 70298.7 |
| 21 | .3972048 | 24.37855 | 70153.99 | 70154.16 | 70303.01 |
| 22 | .3974621 | 22.62832 | 70152.23 | 70152.41 | 70308.03 |
| 23 | .3976145 | 22.00023 | **70151.59** | **70151.79** | 70314.16 |
| 24 | **.3976661** | 22.44891 | 70152.03 | 70152.25 | 70321.38 |
| 25 | .3976087 | 24.06251 | 70153.64 | 70153.88 | 70329.76 |
| 26 | .39752 | 26.01011 | 70155.59 | 70155.84 | 70338.48 |
| 27 | .3974273 | 28 | 70157.58 | 70157.85 | 70347.25 |

predictors for each model:

Figur 7.1: Prediktorer fra VSelect-kommandoen.

Hvorav de ulike modellene er vist i figur 7.2.

```
predictors for each model:

1  :  age
2  :  early_ability age
3  :  early_ability age books_home
4  :  early_ability age par_edu books_home
5  :  early_ability age par_edu not_born books_home
6  :  early_ability age par_edu not_born books_home sameteacher4_plus
7  :  early_ability age par_edu not_born books_home girl sameteacher4_plus
8  :  early_ability birthy birthm age par_edu books_home teacher_edu idgrade
9  :  early_ability birthy birthm age par_edu not_born books_home teacher_edu idgrade
10 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_edu idgrade
11 :  early_ability birthy birthm age par_edu not_born books_home girl income teacher_edu idgrade
12 :  early_ability birthy birthm age par_edu not_born books_home girl speak_testlang_home income teacher_edu idgrade
13 :  early_ability birthy birthm age par_edu not_born books_home girl speak_testlang_home income pc_class teacher_edu idgrade
14 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age speak_testlang_home income pc_class teacher_edu idgrade
15 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income pc_class teacher_edu idgrade
16 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize pc_class teacher_edu idgrade
17 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class teacher_edu
18 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
19 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
20 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
21 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
22 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
23 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
> grade
24 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
> cher_edu idgrade
25 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
> _not_born teacher_edu idgrade
26 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
> _not_born par_emp teacher_edu idgrade
27 :  early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad
> _not_born par_emp schoolsize4 teacher_edu idgrade
```

Figur 7.2: Ulike modeller fra VSelect med ulikt antall variable.

Vselect foreslår følgende modell med 24 variabler med lavest adjusted $R^2$-score: *early_ability birthy birthm age par_edu not_born books_home girl teacher_age school_location speak_testlang_home income clsize teacher_exp pc_class pct_abroad teacher_fem teacher_cert kinderg_att sameteacher4_plus sameteacher1_less pct_disadv idgrade teacher_edu*.

### 7.1.2 Deskriptiv statistikk for kontrollvariabler

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| early_abil~y | 22,558 | 2.619292 | .9636862 | 1 | 4 |
| birthy | 24,704 | 3.362411 | .5950971 | 1 | 7 |
| birthm | 24,719 | 6.375177 | 3.383698 | 1 | 12 |
| age | 24,938 | 10.47091 | .6224268 | 6.667 | 14.583 |
| par_edu | 22,624 | 2.163234 | .998522 | 1 | 5 |
| not_born | 24,220 | .1543765 | .3613168 | 0 | 1 |
| books_home | 23,086 | 3.634454 | 1.264845 | 1 | 5 |
| girl | 24,719 | .4909584 | .4999284 | 0 | 1 |
| school_loc~n | 18,641 | 2.432917 | .9234186 | 1 | 4 |
| speak_test~e | 24,412 | 1.14624 | .4060057 | 1 | 3 |
| income | 15,313 | 3.580748 | 1.685934 | 1 | 6 |
| clsize | 24,082 | 24.13886 | 7.45898 | 2 | 58 |
| pc_class | 23,604 | .6194713 | .4855271 | 0 | 1 |
| pct_abroad | 22,638 | 1.166357 | .5204438 | 1 | 4 |
| kinderg_att | 20,030 | .8668997 | .3396916 | 0 | 1 |
| pct_disadv | 22,613 | 1.961261 | 1.038999 | 1 | 4 |
| idgrade | 24,979 | 3.895392 | .3060536 | 3 | 4 |

Figur 7.3: Deskriptiv statistikk for mulige kontrollvariabler.

## 7.2 VIF-test

| Variable | VIF | 1/VIF |
|---|---|---|
| age | 203.23 | 0.004920 |
| birthy | 161.08 | 0.006208 |
| birthm | 64.59 | 0.015483 |
| teacher_age | 3.08 | 0.324650 |
| teacher_exp | 3.07 | 0.325809 |
| sameteach~us | 1.70 | 0.587445 |
| pct_disadv | 1.70 | 0.589229 |
| pct_abroad | 1.60 | 0.625616 |
| books_home | 1.58 | 0.631116 |
| income | 1.54 | 0.650213 |
| par_not_born | 1.45 | 0.688061 |
| school_loc~n | 1.45 | 0.691855 |
| par_edu | 1.40 | 0.715573 |
| schoolsize4 | 1.38 | 0.726122 |
| pc_class | 1.35 | 0.738931 |
| speak_test~e | 1.33 | 0.749318 |
| teacher_cert | 1.25 | 0.799574 |
| par_emp | 1.25 | 0.802842 |
| clsize | 1.13 | 0.885833 |
| not_born | 1.11 | 0.897931 |
| early_abil~y | 1.10 | 0.911863 |
| sameteach~ss | 1.09 | 0.918219 |
| teacher_fem | 1.04 | 0.957573 |
| kinderg_att | 1.04 | 0.959122 |
| girl | 1.04 | 0.964918 |
| Mean VIF | 18.46 | |

Figur 7.4: Resultatene fra en VIF-test gjennomført i Stata. Testen er gjort på en regresjonsmodell med leseferdighetene som avhengig variabel og alle de andre relevant variablene fra datasettet som forklaringsvariabler.

## 7.3 Chow-test

```
( 1)  sovjet = 0
( 2)  teacher_exp_sovjet = 0
( 3)  teacher_age_sovjet = 0
( 4)  teacher_fem_sovjet = 0
( 5)  teacher_cert_sovjet = 0
( 6)  sameteacher1_less_sovjet = 0
( 7)  sameteacher4_plus_sovjet = 0

      F(  7, 18361) =   20.94
           Prob > F =    0.0000
```

Figur 7.5: Resultater fra Chow-test gjennomført i Stata. Testen er gjort på to regresjonsmodeller med leseferdighetene som avhengig variabel og alle de andre relevante variabler fra datasettet som forklaringsvariabler. Strukturelle forskjeller i dataen er funnet for alle egenskaper for lærere fra Sovjet versus referansegruppen.

## 7.4 Regresjonsanalyser

| VARIABLES | (1) read | (2) read | (3) read | (4) read | (5) read | (6) read |
|---|---|---|---|---|---|---|
| 2.teacher_age | 21.92 | | | | | |
| | (0) | | | | | |
| 3.teacher_age | 21.35 | | | | | |
| | (0) | | | | | |
| 4.teacher_age | 19.08 | | | | | |
| | (0) | | | | | |
| 5.teacher_age | 21.22 | | | | | |
| | (0) | | | | | |
| 6.teacher_age | 26.33 | | | | | |
| | (0) | | | | | |
| early_ability | 21.87 | 22.02 | 22.09 | 21.90 | 22.15 | 21.87 |
| | (0) | (0) | (0) | (0) | (0) | (0) |
| age | 17.64 | 15.63 | 15.95 | 17.10 | 15.99 | 17.89 |
| | (0) | (0) | (0) | (0) | (0) | (0) |
| par_edu | -17.03 | | | | | |
| | (0) | | | | | |
| girl | 13.56 | 13.19 | 13.23 | 13.27 | 13.38 | 13.64 |
| | (0) | (0) | (0) | (0) | (0) | (0) |
| speak_testlang_home | -20.50 | -22.07 | -21.87 | -20.89 | -21.81 | -20.89 |
| | (0) | (0) | (0) | (0) | (0) | (0) |
| clsize | 0.547 | 0.562 | 0.599 | 0.582 | 0.608 | 0.567 |
| | (0) | (0) | (0) | (0) | (0) | (0) |
| pc_class | 13.20 | 15.69 | 15.69 | 14.13 | 15.19 | 10.89 |
| | (0) | (0) | (0) | (0) | (0) | (0) |
| pct_abroad | -0.660 | -1.623 | -1.700 | -1.852 | -1.861 | -3.423 |
| | (0.482) | (0.0813) | (0.0674) | (0.0454) | (0.0454) | (0.000253) |
| teacher_exp | | 0.0988 | | | | |
| | | (0.0154) | | | | |
| higher_edu | | 33.17 | 33.28 | 31.91 | 33.26 | 32.51 |
| | | (0) | (0) | (0) | (0) | (0) |
| teacher_fem | | | 1.977 | | | |
| | | | (0.181) | | | |
| teacher_cert | | | | 16.71 | | |
| | | | | (0) | | |
| sameteacher1_less | | | | | 11.69 | |
| | | | | | (0.0132) | |
| sameteacher4_plus | | | | | | -12.76 |
| | | | | | | (0) |
| Constant | 305.0 | 289.8 | 285.0 | 261.6 | 286.2 | 276.3 |
| | (0) | (0) | (0) | (0) | (0) | (0) |
| Observations | 18,352 | 18,418 | 18,813 | 18,691 | 18,731 | 18,731 |
| R-squared | 0.263 | 0.252 | 0.253 | 0.259 | 0.254 | 0.260 |

p-value in parentheses

Tabell 7.1: Regresjonsanalyse med alle interessevariabler og kontrollvariabler.

| VARIABLES | (1)<br>read | (2)<br>read |
|---|---|---|
| 1.sovjet | -8.293 | -15.46 |
| | (0.0857) | (0.00172) |
| 2.teacher_age | 42.36 | 24.87 |
| | (0) | (8.69e-10) |
| 3.teacher_age | 32.07 | 21.90 |
| | (0) | (2.14e-08) |
| 4.teacher_age | 25.91 | 15.01 |
| | (0) | (0.000111) |
| 5.teacher_age | 40.98 | 22.13 |
| | (0) | (1.07e-08) |
| 6.teacher_age | 25.15 | 15.82 |
| | (1.94e-07) | (0.000785) |
| 1.sovjet#2.teacher_age | -20.68 | -3.586 |
| | (0.000169) | (0.508) |
| 1.sovjet#3.teacher_age | 0.498 | 3.419 |
| | (0.922) | (0.495) |
| 1.sovjet#4.teacher_age | 3.669 | 12.76 |
| | (0.477) | (0.0117) |
| 1.sovjet#5.teacher_age | -12.08 | -2.197 |
| | (0.0208) | (0.668) |
| 1.sovjet#6.teacher_age | 28.08 | 26.45 |
| | (1.08e-05) | (1.93e-05) |
| early_ability | | 21.93 |
| | | (0) |
| age | | 17.54 |
| | | (0) |
| higher_edu | | 32.81 |
| | | (0) |
| girl | | 13.38 |
| | | (0) |
| speak_testlang_home | | -20.61 |
| | | (0) |
| clsize | | 0.586 |
| | | (0) |
| pc_class | | 6.877 |
| | | (1.55e-06) |
| pct_abroad | | -2.472 |
| | | (0.00854) |
| Constant | 502.2 | 261.0 |
| | (0) | (0) |
| Observations | 24,307 | 18,803 |
| R-squared | 0.022 | 0.263 |

p-value in parentheses

Tabell 7.2: Regresjonsanalyser med interaksjonsvariabel for *sovjet* og *teacher_age*.

| VARIABLES | (1)<br>read | (2)<br>read | (3)<br>read | (4)<br>read |
|---|---|---|---|---|
| 1.sovjet | -21.10 | -18.27 | -30.80 | -22.73 |
| | (0) | (0) | (0) | (2.84e-09) |
| teacher_exp | 0.201 | -0.0182 | | |
| | (0.000298) | (0.730) | | |
| 1.sovjet#c.teacher_exp | 0.371 | 0.351 | | |
| | (1.38e-05) | (2.16e-05) | | |
| early_ability | | 21.94 | | 22.00 |
| | | (0) | | (0) |
| age | | 17.44 | | 17.43 |
| | | (0) | | (0) |
| higher_edu | | 33.05 | | 33.16 |
| | | (0) | | (0) |
| girl | | 13.33 | | 13.44 |
| | | (0) | | (0) |
| speak_testlang_home | | -21.29 | | -21.29 |
| | | (0) | | (0) |
| clsize | | 0.585 | | 0.632 |
| | | (0) | | (0) |
| pc_class | | 6.864 | | 7.880 |
| | | (1.85e-06) | | (3.15e-08) |
| pct_abroad | | -2.975 | | -2.920 |
| | | (0.00159) | | (0.00191) |
| 1.teacher_fem | | | -2.101 | 0.780 |
| | | | (0.227) | (0.636) |
| 1.sovjet#1.teacher_fem | | | 19.73 | 12.75 |
| | | | (7.99e-07) | (0.000722) |
| Constant | 532.6 | 283.5 | 537.1 | 280.0 |
| | (0) | (0) | (0) | (0) |
| Observations | 23,679 | 18,418 | 24,321 | 18,813 |
| R-squared | 0.012 | 0.255 | 0.008 | 0.256 |

p-value in parentheses

Tabell 7.3: Regresjonsanalyser med interaksjonsvariabel for *sovjet* og *teacher_exp*, samt *sovjet* og *teacher_fem* .

| VARIABLES | (1)<br>read | (2)<br>read |
|---|---|---|
| 1.sovjet | -37.66 | -32.56 |
| | (0) | (0) |
| 1.teacher_cert | 4.037 | -2.311 |
| | (0.105) | (0.358) |
| 1.sovjet#1.teacher_cert | 33.95 | 26.66 |
| | (0) | (0) |
| early_ability | | 21.72 |
| | | (0) |
| age | | 18.45 |
| | | (0) |
| higher_edu | | 31.42 |
| | | (0) |
| girl | | 13.39 |
| | | (0) |
| speak_testlang_home | | -20.26 |
| | | (0) |
| clsize | | 0.650 |
| | | (0) |
| pc_class | | 7.989 |
| | | (1.75e-08) |
| pct_abroad | | -3.165 |
| | | (0.000729) |
| Constant | 531.5 | 272.6 |
| | (0) | (0) |
| Observations | 24,136 | 18,691 |
| R-squared | 0.028 | 0.263 |

p-value in parentheses

Tabell 7.4: Regresjonsanalyser med interaksjonsvariabel for *sovjet* og *teacher_cert*.

| VARIABLES | (1)<br>read | (2)<br>read | (3)<br>read | (4)<br>read |
|---|---|---|---|---|
| 1.sovjet | -13.04 | -10.31 | -10.65 | -10.54 |
| | (0) | (0) | (0) | (9.02e-10) |
| 1.sameteacher1_less | 35.50 | 29.25 | | |
| | (1.77e-08) | (7.61e-06) | | |
| 1.sovjet#1.sameteacher1_less | -39.49 | -35.26 | | |
| | (1.92e-05) | (0.000185) | | |
| early_ability | | 22.10 | | 22.05 |
| | | (0) | | (0) |
| age | | 17.53 | | 16.98 |
| | | (0) | | (0) |
| higher_edu | | 33.25 | | 32.87 |
| | | (0) | | (0) |
| girl | | 13.52 | | 13.59 |
| | | (0) | | (0) |
| speak_testlang_home | | -21.38 | | -21.06 |
| | | (0) | | (0) |
| clsize | | 0.625 | | 0.570 |
| | | (0) | | (0) |
| pc_class | | 7.392 | | 7.086 |
| | | (1.93e-07) | | (5.62e-07) |
| pct_abroad | | -3.207 | | -3.958 |
| | | (0.000671) | | (2.61e-05) |
| 1.sameteacher4_plus | | | -30.71 | -20.38 |
| | | | (0) | (0) |
| 1.sovjet#1.sameteacher4_plus | | | 20.01 | 15.56 |
| | | | (0) | (0) |
| Constant | 535.7 | 280.0 | 540.3 | 291.2 |
| | (0) | (0) | (0) | (0) |
| Observations | 24,069 | 18,731 | 24,069 | 18,731 |
| R-squared | 0.010 | 0.257 | 0.023 | 0.262 |

p-value in parentheses

Tabell 7.5: Regresjonsanalyser med interaksjonsvariabel for *sovjet* og *sameteacher4_plus*, samt *sovjet* og *sameteacher1_less*.

## 7.5 Mønstergjenkjenning

Følgende kode ble brukt for å finne mønstre for lærere. Programvarebibliotekene Pandas og Mlxtend ble brukt til å gjennomføre denne delen av analysen. Gjenbruk gjerne koden til videre undersøkelser av temaet.

```
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules


df=pd.read_stata("bachlor_dataset regresjonsvariabler.dta")
df1=df[["read", "sovjet", "teacher_age", "teacher_cert", "
                                    sameteacher1_less"]]


def bucketizingdata(df): #Changing values of read to number 1-4
                                    depending of which 25% they relate
                                    to.
    newdf = df.copy()
    for col in newdf:
        if col not in ["teacher_age", "teacher_cert","sameteacher1_less"
                                                , "sovjet"]:
            newdf[col] = pd.qcut(newdf[col], 4, labels=['1', '2', '3', '
                                                4'])
    newdf=newdf[newdf.read == '4']
    print(newdf)
    return newdf


def dummifyandapriori(df): #Dummifies all values used and uses the
                                    apriori algorithm to generate a list
                                     of frequent itemsets, which are
                                    then filtered by parameters of
                                    choice
    dummylist = []
    for att in df:
        if att in ["read", "teacher_age", "teacher_cert","
                                            sameteacher1_less", "sovjet"
                                            ]:
            df[att] = df[att].astype("category")
        dummylist.append(pd.get_dummies(df[[att]]))
    dummified_df = pd.concat(dummylist, axis=1)

    frequent_itemsets = {}
    minpaterns = 300
    minsup = 1.0
    minconf = 0.9
    while minsup > 0:
        minsup = minsup * 0.95
        frequent_itemsets = apriori(dummified_df, min_support=minsup,
                                            use_colnames=True)
        if len(frequent_itemsets) >= minpaterns:
            print("The perfect teacher:")
            print("Minimum support:", minsup)
            break
    print("Number of patterns:", len(frequent_itemsets))
    rules = association_rules(frequent_itemsets, metric="confidence",
```

```
                                        min_threshold=minconf)
    rules["antecedent_len"] = rules["antecedents"].apply(lambda x: len(x
                                        ))
    rules["consequents_len"] = rules["consequents"].apply(lambda x: len(
                                        x))
    newrules = rules[(rules['antecedent_len'] >= 1)]
    newrules = newrules[(rules['consequents_len'] == 1)]
    #newrules = newrules[(newrules['consequents']=='read_10')]
    if len(newrules) < 1:
        print("No rules found with given threshold")
    sortednewrules = newrules.sort_values("lift", ascending=False)
    score_lift = mean_lift(sortednewrules)
    score_confidence = mean_confidence(sortednewrules)
    print("Mean lift:", score_lift)
    print("Mean confidence:", score_confidence)
    return sortednewrules


def top_rules(rules): #selects the 100 top rules
    top_df=rules.head(100)
    return top_df


def patternmining_teacher(df):
    a= dummifyandapriori(bucketizingdata(df))
    return a


def main(): #Connects all the tasks and saves the file as a CSV
    top_rules(patternmining_teacher(df1)).to_csv('YOURFILENAME.csv',
                                        index=True,header=True)


if __name__=="__main__":
    main()


def mean_lift(result):
    metric = "lift"
    measure = result.loc[:, metric]
    measure_sum = 0
    for m in measure:
        measure_sum = measure_sum + m
    if len(measure) == 0:
        mean_measure = 0
    else:
        mean_measure = measure_sum / len(measure)
    return mean_measure


def mean_confidence(result):
    metric = "confidence"
    measure = result.loc[:, metric]
    measure_sum = 0
    for m in measure:
        measure_sum = measure_sum + m
    if len(measure) == 0:
        mean_measure = 0
```

```
    else:
        mean_measure = measure_sum / len(measure)
    return mean_measure
```

# Referanser


Barth, E., 2005, Samfunnsmessig avkastning av utdanning, ISF (https://www.ssb.no/a/publikasjoner/pdf/sa74/kap-8.pdf, Sist lastet ned: 03.05.2020)

Bonesrønning, H., 2008, Prestasjonsforskjeller mellom skoler og kommuner: Analyse av nasjonale prøver - SØF rapport nr. 01/10 ( http://www.sof.ntnu.no/SOF-R%2001$_1$0.*pdf*, Sist lastet ned: 03.05.2020)

Borgelt, C., Kruse, R., 2002, Induction of association rules: Apriori implementation, Springer

Carlsen, F., 2015, Kronikk; Fra kommunisme til kapitalisme (https://www.nrk.no/ytring/fra-kommunisme-til-kapitalisme-1.12612489, Sist lastet ned: 03.05.2020)

FN, 2020a, Moldova, (https://www.fn.no/Land/Moldova, Sist lastet ned: 03.05.2020)

FN, 2020b, Lativa, (https://www.fn.no/Land/Latvia, Sist lastet ned: 03.05.2020)

Furnival, G.M., Wilson, R.W., 1974, Regressions by leaps and bounds, Technometrics 16(4): 499-511

Gornitzka, Å., 2003, Kvalitet i norsk høyere utdanning i et internasjonalt perspektiv - En del-utredning for Ryssdalutvalget, NIFU skriftserie nr. 25/2003, (https://nifu.brage.unit.no/nifu-xmlui/bitstream/handle/11250/280520/NIFUskriftserie2003-25.pdf?sequence=1, Sist lastet ned: 03.05.2020)

Gould, W., 2014, How to Compute the Chow test Statistic? (https://www.stata.com/support/faqs/statisti
chow-statistic/, Sist lastet ned: 13.05.2020)

Hanushek, E., 2020, Education production functions (http://hanushek.stanford.edu/sites/default/files/publications /Hanushek%202020%20Education%20Production%20Functions.pdf, Sist lastet ned: 03.05.2020)

IEA, 2016b, Summary of international results (http://timssandpirls.bc.edu/pirls2016/international-results/pirls/summary/ and http://timssandpirls.bc.edu/pirls2016/international-results/wp-content/uploads/structure/PIRLS/P16-International-Findings-from-PIRLS-2016.pdf, Sist lastet ned: 03.05.2020)

IEA, 2016a, PIRLS results (http://timssandpirls.bc.edu/pirls2016/international-results/wp-content/uploads/structure/PIRLS/0.-about-pirls-2016/0_1_pirls-2016-countries.pdf, Sist lastet ned: 03.05.2020)

Jelstad, J., 2015, Et overveldende antall studier viser ingen effekt av økt lærertetthet (https://www.utdanningsnytt.no/laerertetthet/et-overveldende-antall-studier-viser-ingen-effekt-av-okt-laerertetthet/188678, Sist lastet ned: 03.05.2020)

MathWorks, 2020, Adjusted $R^2$ sscore (https://www.mathworks.com/help/matlab/data_analysis/linear-regression.htmlbswinlz, Sist lastet ned: 12.05.2020)

Regjeringen, 2019, NOU 2019: 2 Fremtidige kompetansebehov II — Utfordringer for kompetansepolitikken (https://www.regjeringen.no/no/dokumenter/nou-2019-2/id2627309/?ch=9, Sist lastet ned: 03.05.2020)

Thomas, R. L., 2005, Using Statistics in Economics, McGrawHill

TradingEconomics, 2020, Bruttonasjonalprodukt (https://tradingeconomics.com/russia/gdp,
https://tradingeconomics.com/lithuania/gdp,
https://tradingeconomics.com/moldova/gdp,
https://tradingeconomics.com/norway/gdp,
https://tradingeconomics.com/sweden/gdp,
https://tradingeconomics.com/iceland/gdp,
Sist lastet ned: 03.05.2020)

Utdanningsforbundet, 2017, (https://www.utdanningsforbundet.no/var-politikk/stopp-store-klasser/forskning-og-fakta/stotte-i-forskningen/, Sist lastet ned: 03.05.2020)

Wold, S., Esbensen, K., Geladi, P., 1987, Principal component analysis, Chemometrics and intelligent laboratory systems, 2(1-3):46-47

Worldometers, 2020, Befolkningstall, (https://www.worldometers.info/world-population/population-by-country/, Sist lastet ned: 03.05.2020)

NTNU

Kunnskap for en bedre verden