UiT Norges arktiske universitet

Handelshøgskolen ved UiT

Kan karaktersnitt i matematiske fag brukes til å predikere antall søkere til IT-studier?

Martin Nilsen Bless

Masteroppgave i samfunnsøkonomi, SOK-3901, juni 2020

Forord

Takk til Espen Sirnes som har vært min veileder på denne oppgaven.

Jeg vil også takke mine foreldre fordi de er så kule.

Og sist, men ikke minst vil jeg takke de jeg har studert sammen med, for interessante samtaler (AKA distraksjoner), intense runder med bordtennis og så voldsomt mye pant.

Abstrakt

Digitalisering blir en større del av hverdagen. Som følge av dette blir etterspørselen etter arbeidskraft innenfor informasjonsteknologi også større. For å tilby arbeidskraft er utdanning viktig. Oppgaven skal vise hvilken funksjon utdanning har i tilbud av arbeidskraft, ved å se på karaktersnitt og hvordan dette påvirker søkertall. Mer spesifikt skal oppgaven se på hvordan karaktersnitt i matematikk fagene R2 og S2 påvirker hvor mange som søker på informasjonsteknologi studie på universitetet. Dette gjøres ved å skape lineære modeller basert på karakterdata og søkertall, og sette de opp imot hverandre. Modellene testes så for signifikans, og endres basert på resultater. Resultatene i fra oppgaven er at karaktersnitt i R2 er signifikante for søkere innen IT, kvalifiserte søkere innen IT og førstevalgs søkere inne IT, der signifikansen er størst for førstevalgs søkere. Karaktersnitt i S2 er derimot ikke signifikant for noen av søkertallene. Resultatene kommer frem fra data med få målinger, som setter resultatene under tvil. Det som er sikkert er at hvis man skulle bruke gjennomsnittlige matematikk karakterer til å predikere søkertall på IT-studier, ville det vært bedre å bruke karakterer ifra R2 enn ifra S2.

Alle beregninger gjennomført i oppgaven er gjort ved bruk av koder i R-studio.

Nøkkelord: IT, IKT, lineær regresjon, humankapital, utdanning

Innhold

Tabelliste

Figurliste

1 Introduksjon

Digitalisering blir en større og større del av samfunnet. Ikke bare i hverdagen, men også i arbeidslivet. Digitalisering er ansett som en av de største trendene som påvirker samfunn og business (Parviainen et al., 2017). Ifølge Parviainen et al. er en omfattende del av digitalisering «forandring assosiert med applikasjon av digital teknologi i alle aspekter av mennesket sitt samfunn». Naturligvis har det også en effekt på jobbene i bedrifter. I en mer digital hverdag etterspørres gjerne også mer digital kunnskap.

For å bruke digital teknologi kreves kunnskaper innen relevante felter. Et av disse er IT. IT er informasjons *teknologi*, så en økning i digitalisering fører til en økning i bruk av IT. Innenfor IT finner man IKT, som er informasjons og kommunikasjons teknologi. En økning i bruk av IT gir en økning i bruk av IKT. IKT er definert som enten digitale verktøy eller digitale infrastrukturer som bidrar til overføring av informasjon (Zuppo, 2012). Forskning innen dette feltet gir resultater som påvirker både IKT og IT. Det er viktig å forstå sammenhengen mellom de to, og at de påvirker hverandre veldig direkte. I oppgaven kommer det til å antas at all forskning innen IKT påvirker IT.

En ting som trekkes frem som viktig innen utdanning til IKT er matematikk kunnskaper. Dette ser man både i forskning i fra Sør-Afrika som omhandler mangel på IKT-kompetanse (de Villiers et al., 2012), og forskning i fra Estland som ser på hvordan man kan bruke matematikk-egenskaper til å predikere om noen droppet ut av studie eller ikke (Kori et al., 2015). Begge artiklene gjenspeilet det faktum at matematikk egenskaper har en viss effekt på utdanning innenfor IKT.

En annen grunn til at man kan hevde at matematikk er viktig for utdanning innen informasjonsteknologi (heretter referert til som IT-studier) er opptakskrav til IT-studier på flere universiteter i Norge. Dette gjelder Universitet i Tromsø (UiT Norges Arktiske Universitet, 2022), Universitet i Oslo (Universitetet i Oslo, 2022) og Norges teknisk-naturvitenskaplige universitet (NTNU, 2022). Alle de tre universitetene har spesielle krav for opptak i IT-studier som krever ekstra matematikk-kunnskap, det vil si krav om enten S1 og S2 matematikk, eller R1 matematikk.

Ut ifra dette kommer hensikten med denne oppgaven. Hvis universiteter og tidligere forskning tilsier at man må ha en viss mengde matematisk kunnskap for å studere informasjonsteknologi, kan man da også bruke kunnskapsnivå som en pekepinn på hvor

mange som kommer til å søke på denne typen studie? Med andre ord er hensikten å teste om andelen søkere (totalt, førstevalg og kvalifiserte) til IT-studier varierer når ferdighetsnivået innen matematikk varierer. Dette ferdighetsnivået vil kunne måles ved bruk av snitt karakterer i noen matematiske fag nasjonalt ved det tidspunktet når man søker på studie. Tidspunkt når man søker for studie er gjerne (men ikke eksklusivt) etter fullført tredje år videregående skole. For enkelhets skyld brukes kun karaktersnitt ifra matematikk fag fra tredje år videregående.

Selv om opptakskravene er for R1, har oppgaven brukt R2 snitt som mål, ettersom dette representerer nivået ved søketidspunkt. Av samme årsak brukes kun S2, og ikke S1 sine snittkarakterer i oppgaven. Bruken av R2 representerer ferdighetsnivå, men ikke opptakskrav. Dette betyr at resultatet ikke kan si noe om valget av opptakskrav. I tillegg bruker oppgaven også snitt innen faget informasjonsteknologi 2 på videregående skole. Årsaken til dette er for å kunne trekke en logisk beslutning ut av resultatene. Hvis statistiske tester tilsier at ferdighet innen et av de matematiske fagene er en mer nøyaktig måte å predikere søkertall på IT-studier enn ferdighet innen informasjonsteknologi-faget på videregående, kan dette være et tegn på at metoden har mangler i seg.

2 Tidligere forskning

Tidligere forskning i denne oppgaven er sentrert rundt to punkter.

Det ene er forskning relatert til økonomisk teori. Dette er forskning som gjelder generelt i økonomisk teori, og ikke omhandler spesifikt temaene informasjonsteknologi og informasjon- og kompetanseteknologi. Dette involverer definisjoner av noen begreper som dukker opp gjennom oppgaven, og gjennomgang av noen økonomiske modeller.

Det andre er forskning relatert til temaet. Dette vil være rapporter og artikler om forskjellige aspekter ved IT/IKT.

Til sist kommer hypotesen, som baserer seg på tidligere forskning og generell økonomisk teori.

2.1 Økonomisk teori

2.1.1 Solow med humankapital:

Dette avsnittet er hentet ifra (Jones & Vollrath, 2013).

Solow modellen viser hvordan produksjon avhenger av kapital og arbeidskraft. Dette ser vi i likning (1):

$$Y = K^{\alpha}L^{1-\alpha} \qquad (1)$$

Når modellen tar hensyn til humankapital utvides definisjonen av arbeidskraft slik at den nye likningen, likning (2), ser ut som dette.

$$Y = K^{\alpha}(AH)^{1-\alpha} \qquad (2)$$

Solowmodellen ble vurdert til at den presterte veldig bra, men Gregory Mankiw, David Romer og David Weil mente at modellen kunne «passe» bedre ved å utvides til å inkludere humankapital. Ved å inkludere humankapital gjenkjenner man at det kan være forskjellige nivåer av utdanning og ferdigheter i arbeidskraften i forskjellige økonomiske markeder.

Individer i dette nye hypotetiske økonomiske markedet får humankapital ved å bruke tid på å lære skills i stedet for å arbeide. Humankapitalen defineres ved likning (3).

$$H = e^{\psi u}L \qquad (3)$$

H står for skilled labour, ψ angir forventet lønnsvekst av 1 år ekstra utdanning, u er tid brukt på å lære nye skills, og L er total mengde labour som brukes i produksjonen i et økonomisk marked. Intuitivt, hvis u = 0 vil det si at ingen andel av labour er skilled (H = L), ettersom $e^0 = 1$. H øker når u øker. Det vil si at produktivitet i økonomien øker når individer bruker mer tid på å lære nye skills, og følgelig øker produktiviteten til individet ved akkumulering av skills også. Det ser vi på likningen, likning (4), for output per arbeider.

$$y = \left(\frac{s_K}{n + g + \delta}\right)^{\frac{\alpha}{1-\alpha}} hA \quad (4)$$

Den deriverte av y med hensyn på h er positiv, det vil si at en økning i h vil føre til en økning i y. Vi ser dette på hvilke verdier vi forventer ifra de forskjellige variablene. (n + g + δ) forventes å være positiv. Sk har samme forventninger, siden man ikke kan spare negativt, og det A kan heller ikke være negativ. Eksponenten $\frac{\alpha}{1-\alpha}$ vil heller aldri føre til at likning (4) blir negativ, siden α er større enn 0 og mindre enn 1. Den deriverte ser vi i likning (5).

$$\frac{dy}{dh} = \left(\frac{s_K}{n + g + \delta}\right)^{\frac{\alpha}{1-\alpha}} A \quad (5)$$

Følgelig er den derivert av y med hensyn på h positiv.

Likning (3) for humankapital i arbeids-styrken kan forenkles til en likning for den individuelle arbeider h.

$$H = e^{\psi u} L \rightarrow h = e^{\psi u} \quad (6)$$

Siden ψ er et tall mellom 0 og 1, og u er positiv vil eksponenten alltid være positiv. e opphøyd i et positivt tall vil alltid være positiv. Derav er den deriverte av h med hensyn til u positiv.

Ettersom det antas at produktivitet gir høyere lønn tilsier dette at det er profitabelt å øke u. Ut ifra dette kan vi si at et individ får en fortjeneste av å utdanne seg. Hvor mye man skal utdanne seg avhenger av kostnader ved utdanningen. Det kommer frem i neste kapittel.

2.1.2 Humankapital og utdanning:

Dette avsnittet er hentet ifra (Cahuc et al., 2014).

Becker (1964) begynner med en hypotese om at utdanning er en investering som vil produsere fremtidig gevinst. Dette forutsetter at lønning blir fastsatt etter produktivitet, og at produktivitet igjen blir påvirket av utdanning gjennom livet. Investeringen kommer med en kostnad. Disse kostnadene inkluderer studiekostnader, alternativkostnad i form av tapt inntekt siden man velger utdanning i stedet for å jobbe, og psykologiske kostnader i form av stress og motgang. Den samlede kompetansen etter utdanningen, som øker produktiviteten til individet, kalles humankapital.

Ifølge Becker kan utdanning bare lede til fremtidig gevinst når produktivitet dikterer lønn. Det er ikke åpenbart at økt produktivitet leder til økt lønn. En årsak er at hvis arbeideren kun har én jobbmulighet, vil ikke lønnsgiver ha insentiver til å øke lønna. Derimot, hvis det er flere alternativer kan arbeideren true med å skifte jobb til en arbeidsgiver som gir bedre lønn. Becker skiller mellom to typer utdanning (training). Disse er generell training og spesifikk training. Generell training øker en arbeider sin produktivitet for alle typer jobb, mens spesifikk training bare øker produktivitet for en spesifikk jobb.

IT-studie er et eksempel på spesifikk training, der utdanningen ikke nødvendigvis leder til økt produktivitet overalt i arbeidsmarkedet, men det kan definitivt bidra til økt produktivitet innen IT-relaterte jobber.

Hvis man skal utdanne seg må man være tilstrekkelig tålmodig, og avkastning av utdanning må være tilstrekkelig høy.

Boka definerer et tidspunkt s, der marginal avkastning er positiv frem til dato s, og negativ etter dato s. Det betyr at det er i individet sin interesse å bruke all sin tid til utdanning før dato s, og ingen tid etter dato s. Dato s uttrykkes med likning (7) og likning (8).

$$s = T + \frac{1}{r}\ln\left(\frac{\theta - r}{\theta}\right) \; hvis \; \theta \geq \frac{r}{1 - e^{-rT}} \qquad (7)$$

$$s = 0 \; hvis \; ikke \; \theta \geq \frac{r}{1 - e^{-rT}} \qquad (8)$$

Hvor lenge man utdanner seg øker med levetid og effektiviteten til individet. De som er mest effektive burde utdanne seg lengere. Lengden av utdanning faller når r stiger, som angir

diskonteringsraten. Det betyr at utålmodige individer, og de som møter høyest finansielle utfordringer bak utdanning sin (og dermed må ta høyere lån/opplever høyere kostnader), burde studere i kortere perioder. I Norge er levetiden høyere enn EU gjennomsnitt (Ringard et al., 2013), og kostnaden av å studere er lav (Naess, 2020). Det er rimelig å anta at s-datoen i Norge som følge av dette er høy.

Om noen ønsker å søke videre utdanning er s datoen relevant. Det er ikke noe man som individ har informasjon om på søke-tidspunktet, men teoretisk gir det et uttrykk på optimal utdannings-tid. Ved en lav s, kan man muligens ikke tjene på videreutdanning i det hele tatt.

2.1.3 Økonomiske begreper

Dette avsnittet blir en gjennomgang av flere begreper som dukker opp i den tematiske forskningen. Begrepene blir satt i sammenheng ved IT og IKT. Disse begrepene er ikke direkte relevante for metoden og testen som gjennomføres, men brukes for å få forståelse av forskjellige aspekter ved utdanning av IT-spesialister, og tilbud av denne formen for arbeidskraft, som er temaet i oppgaven.

Til å begynne med er det begrepet humankapital. I Goldin defineres humankapital som skills som arbeidsstyrken har, og anses som en ressurs(Goldin, 2016). Det finnes investeringer i personer, for eksempel i form av utdanning. Disse investeringene øker et individ sin produktivitet. Den eneste måten for store mengder vanlige folk å få investert i seg selv er via skole.

Det er et kappløp mellom tilbud av skills og etterspørsel etter skills, med avkastningen av utdanning som en måte å nå likevekt. Når avkastningen er høy vil tilbud av nye skills bli større, mens når avkastningen er lav vil tilbud av nye skills være mindre. Teknologisk fremgang gjennom det siste århundre øker etterspørselen for mer humankapital (Goldin, 2016). Dette kan for eksempel ta form som etterspørsel etter ny skills, som kompetanse innen informasjonsteknologi.

Et annet begrep som dukker opp, er brain drain. Det referer til migrasjonen av individer med høyt skill-nivå (Docquier & Rapoport, 2012). Dette er relevant, da skill som forlater et land kan være et problem for noen land etter fullført utdanning. Konsekvensen av dette kan være at man får en labour shortage. Dette er en stor mengde ledige stillinger som ikke blir fylt opp på grunnlag av en mangel i kvalifiserte arbeidere, som implisert i Cappelli (Cappelli, 2005). Det

er ikke spesifisert hvorvidt det er relevant om det er generalister eller spesialister. Det spiller trolig ingen rolle for definisjonen sin del.

Når det er mange ledige stillinger, påvirker det market tightness i arbeidsmarkedet. Dette er en betegnelse for balansen mellom etterspørsel for, og tilbud av, arbeidskraft (Brigden & Thomas, 2003). Stramheten (tightness) indikerer konkurranse for hver jobb. Når det er mange ledige jobber og få søkere er markedet strammere (López-Bassols, 2002).

Skill shortage, eller mangel på skill, er når arbeidsgivere har vanskeligheter med å fylle ledige jobber for et yrke (Richardson, 2007). Eksempel på dette kan være mangel av spesialisert skill innen et yrke, for eksempel at IKT-spesialister som yrke mangler folk med kompetanse innen IKT-sikkerhet. Et skill gap derimot, oppstår blant de som allerede er ansatt. Skill gap er når arbeidere ikke har de nødvendige kvalifikasjonene, erfaringene eller spesialisert nok skill for å møte firmaet sin etterspørsel (Richardson, 2007). Dette spiller inn når det kommer til EVU, det vil si etter-og-videre utdanning, som foreslås som en løsning for å fylle kompetansegapet innen IKT-sikkerhet (Mark et al., 2017), som er en spesialisering innen IKT.

2.2 Litteratur

Dette kapitlet vil gå gjennom forskning på IT og IKT som er relevant for oppgaven. Det er til sammen ni rapporter og forskningsartikler. De omhandler forskjellige aspekter ved IT relaterte yrker (som også inkluderer IKT) både innenlands og utenlands. Fokuset i oppgaven ligger på tilbud av kompetanse, men noen av disse artiklene nevner også andre mulige måter for å imøtekomme stor etterspørsel etter arbeidskraft innen IT og IKT. Det vektlegges også forskjellige økonomi-teoretiske aspekter, og begreper som er gjennomgått tidligere i oppgaven dukker opp i disse rapportene og artiklene.

At etterspørselen etter arbeidskraft innen IKT er større enn tilbudet er en gjenganger i flere av rapportene. Ifølge Eggen et al. er det et stort udekket behov for IKT kompetanse i det norske arbeidsmarkedet (Eggen et al., 2021). Dette underbygges av Mark et al. som har undersøkt IKT-sikkerhet spesifikt (Mark et al., 2017). De kommer frem til at det vil være et underskudd på IKT-sikkerhetskompetanse på 4000 arbeidere i 2030. Mer generelt viser de til at IT bransjen hadde 6000 ubesatte stillinger i 2015.

Cedefop, som er et byrå under EU med fokus på yrkesrettet opplæring i EU, identifiserer IKT som en av de største manglene i arbeidsmarkedet (Cedefop, 2016). Tilbud av IKT-kompetanse fra nyutdannede er ifølge dem ikke tilstrekkelig.

Kori et al. har funnet at det er mangel på tilbud av kompetanse i Estland (Kori et al., 2015). Ifølge Kori et al. ville det ikke være et tilbudsproblem hvis alle som studerte fullførte graden sin, men de fant ut at frafallet var stort innen IKT i Estland (så mange som 32.2% droppet ut i løpet av første studieår).

Også i Sør-Afrika er manglende tilbud et problem. Ifølge de Villiers har dette konsekvens for hele det økonomiske markedet, ettersom vekst i økonomien i Sør-Afrika avhenger av IKT (de Villiers et al., 2012).

IKT er identifisert av Cedefop som et av de yrkene med størst mangel av kompetanse i arbeidsmarkedet. Det støttes av McGrath, som identifiserer IKT som et av de fem yrkene med størst mangel på arbeidskraft (McGrath, 2021).

Hvorfor er IKT et yrke med så høy mangel på arbeidskraft? En årsak kan være et andelen jobber vokser raskt. I Canada, Kina og Singapore vokser IKT raskere enn den generelle økonomien i landene (Duell, 2020). En vekst i IKT burde lede til flere IKT-relaterte jobber. I Norge øker sysselsettingen. I perioden 2009-2019 økte sysselsettingen i IKT i Norge med 40% (Eggen et al., 2021). Sysselsettingen i landet generelt var 8%. Dette samfaller med en økning i tilbud. At flere får jobb kan være et tegn på mange ledige jobber, og at etterspørselen etter arbeidskraft er stor. Det kan også være tegn på et stramt arbeidsmarked. Et stramt arbeidsmarked er som sagt når det er mange ledige jobber, og de som er nylig utdannet rask får jobb (López-Bassols, 2002).

Som vi kan se i flere av rapportene og forskningsartiklene er det flere land som opplever en økt etterspørsel etter IKT-kompetanse. Ifølge Duell gjelder dette blant andre Canada, Kina, Singapore og Tyskland (Duell, 2020). Det gjelder også i Norge (Eggen et al., 2021), noe også Mark et al. støtter (Mark et al., 2017). Vi ser dette i Estland også (Kori et al., 2015), faktisk i hele EU foruten Finland (Cedefop, 2016). Til sist er dette problemet også identifisert i Sør-Afrika (de Villiers et al., 2012). I de forskjellige landene identifiseres forskjellige årsaker til problemet. Kori et al. finner at tilbudet ville møtt etterspørselen hvis det ikke var for frafall. Eggen et al. ønsker en økning av studieplasser i Norge. Og i Sør-Afrika kan brain drain identifiseres som et potensielt problem, da utdannede kan fristes til å flytte utenlands til land med bedre velferd.

Brain drain, som er migrasjon av skill til et annet land (Docquier & Rapoport, 2012), er ikke bare identifisert som et mulig problem i Sør-Afrika. Mark et al. nevner brain drain, men skal

man tro Duell er det ikke nødvendigvis et stort problem. I følge Duell er det insentiver på plass for utdannede i mindre utviklede land for å migrere til mer utviklede land, som for eksempel bedre velferd (Duell, 2020). I følge Ringard et al. er velferden i Norge god (Ringard et al., 2013).

En løsning på problemet, som ikke omhandler å øke tilbud av kompetanse er etter-og-videreutdanning (EVU). Nødvendig kompetanse for jobber innen IKT øker når feltet utvikler seg over tid. Men kompetansen til de som allerede er utdannet vil ikke alltid holde følge. EVU er av Mark et al. foreslått som en løsning på problemet med at nødvendig kompetanse og virkelig kompetanse ikke er lik (Mark et al., 2017). Eggen et al. er også enig i at EVU er viktig i et fagfelt som utvikler seg så raskt som IKT (Eggen et al., 2021). De vektlegger også at import av arbeidskraft ikke løser problemet, ettersom andre land opplever mangel på kompetanse også, og at Norge mer enn disse landene stiller krav til at arbeidere skal beherske norsk språk for å arbeide i landet.

Dette kapitlet har gjennomgått flere aspekter av IKT-mangelen globalt og nasjonalt. Fokus i oppgaven kommer til å bli på tilbuds-siden av kompetanse. Det er likevel viktig å vite om brain drain og etter-og-videreutdanning blant annet, for å få et større bilde av situasjonen. Basert på dette kapitlet trekkes den følgende hypotesen:

Hypotese: Matematikk karakterer kan predikere søkertall

Hypotesen tar grunnlag i ideen om en høy s-dato, som kommer av at man har høy leve tid i Norge (høy T), og lave studiekostnader (som fører til en lav r). Den tar også grunnlag i at siden IT/IKT enda har undersjangrer (IKT-sikkerhet for eksempel) er det nok jobbmuligheter til å gi incentiv til videre utdanning. Årsaken er at med flere jobbmuligheter spiller produktivitet en større rolle på lønninger som i teori om humankapital. Man kan anta at de med høyere kompetanse (skill) er mer effektive. For eksempel at de som er flinke i matematiske fag er mer effektive enn de som gjør det dårligere. Per definisjonen av s-datoen kan man ut ifra dette også forvente at karakterer kan reflektere de individene som har høyere s-dato, og derved tjener på å utdanne seg lengere. Følgelig, hvis matematikk er viktig for å ta IT-studier, da burde et høyere gjennomsnitt i matematikk lede til at flere ønsker å ta IT-studier. Rent intuitivt virker det dermed som om matematikk burde være best i å predikere kvalifiserte søkere.

3 Data og metode

Dette kapitlet vil gå gjennom data, hvor det er hentet fra og hva det inneholder, og metode, hva oppgaven tester og hvordan den tester det.

3.1 Data

3.1.1 Hvilke data blir brukt?

I denne oppgaven blir det brukt to datasett som er hentet ifra to forskjellige kilder.

Det ene er hentet fra Samordna opptak (data er mottatt på mail). Dette er tall på hvor mange som søker på spesifikke studielinjer. Oppgaven bruker tallene på studielinjen «informasjonsteknologi», som i datasettet heter INFOTEKN. Tallene som blir brukt i oppgaven omhandler søkere til informasjonsteknologi (heretter henvist til som søkere), kvalifiserte søkere og de som søkte IT-studier som førstevalg.

Den andre kilden som blir brukt er udir.no (Utdannings-direktoratet, 2021), der det er hentet tall på gjennomsnittlige karakterer (heretter referert til som snittkarakter eller bare snitt) i fag på videregående skole. Det relevante data som er hentet er gjennomsnittlige karakterer i de tre fagene matematikk R2, matematikk S2 og Informasjonsteknologi 2. I datasettet refereres disse datapunktene til som R2_snitt, S2_snitt og Infotek_snitt_vgs, respektivt.

Begge datasettene har tall som er målt en gang i året, og over den samme tidsperioden. Tallene som brukes i oppgaven er fra 2009 til 2021. For enkelhets skyld er gjennomsnitt karakterer registrert som 2008-09 referert til som 2009, 2009-10 som 2010 osv. Datasettet fra udir.no inneholdt også tall fra 2008, men disse er ekskludert både på grunnlag av at de ikke kunne settes opp mot data ifra Samordna opptak, og på grunnlag av at det var gjennomsnitt hentet fra veldig få elever (20 elever i matematikk og 23 elever i informasjonsteknologi 2 i 2008) sammenliknet med de resterende årene (informasjonsteknologi 2 hadde 2 525 antall elever i 2009).

3.1.2 Svakheter med data

En svakhet ved datasettet er størrelsen, mer spesifikt mengden målinger. Større andel målinger er bedre (Taherdoost, 2017). Taherdoost referer her til surveys, og ikke til målinger som er observert over tid (timeseries). Likevel virker det trygt å anta at det samme gjelder for timeseries, som er formen på datasettet i denne oppgaven.

I timeseries er det ikke nødvendigvis viktig å ha mange målinger for å kunne trekke beslutninger ut av modeller. Det avhenger litt av situasjonen. I en analyse av bilkrasj hevdes det at desto mer stabilt et land er, desto bedre er det med flere målinger (Hassouna & Al-Sahili, 2020). Bilkrasj er ikke utdanning. Men det kan virke som stabiliteten Hassouna og Al-Sahili referer til kan sammenliknes med varians. Lav varians vil si mindre sprik mellom målingene, og klassifiseres som stabile (Breiman, 1996) . Trolig vil det være mer akseptabelt med færre målinger når varians i de uavhengige variablene er tilstrekkelig høy. Varians i R2_snitt i datasettet er 0.032, S2_snitt er 0.017, S2_snitt er 0.017 og Inoftek_snitt_vgs er 0.027. Tilsynelatende virker det som variansene er veldig lave, og indikerer at datasettet burde hatt flere målinger.

Det er også et interessant tilfelle ved at den totale andelen søkere (alle studier) og den totale andelen førstevalgs søkere (alle studier) ikke er identiske. Totale søkere og totale førstevalgs søkere er funnet ved å summere alle søkere og alle førstevalgs søkere respektivt. At de ikke er identiske vil tilsi at noen har søkt på studier, men ikke har noen studie som førstevalg, som ikke gir helt mening. Det kan også være et tegn på at en studieretning er blitt oversett når totale førstevalgs søkere ble summert. Når dette brukes senere i oppgaven vil totalt antall førstevalgs søkere brukes i tilfelle relatert til førstevalgs søkere til IT-studier. Differansen mellom de to søkertallene er gjerne ikke så stor at det burde spille en stor rolle.

3.2 Metode

Målet med oppgaven er å teste for sammenheng mellom karaktersnitt i matematikk fagene R2 og S2, og søkertall på IT-studier. I tillegg til karaktersnitt i matematikk brukes også karaktersnitt i informasjonsteknologi, som er et eget fag på videregående skole, som variabel. Det er trygt å anta at hvis karakterer har en påvirkning på søkertall, så burde informasjonsteknologi (videregående skole) påvirke informasjonsteknologi (universitet). Resultatene for informasjonsteknologi videregående skole (heretter referert til som infotek snitt) spiller dog ikke en stor rolle i oppgaven, siden hovedfokus er på de spesielle kravene til matematikk på IT-studier, som henvist tidligere i oppgaven blir R2 og S2.

I oppgaven er R-studio benyttet for å lage modeller, og til å gjennomføre eventuelle tester av modellene. R-kodene vil finnes i appendiks senere i oppgaven.

3.2.1 Gjennomgang av metoden

For å teste for sammenheng virker det mest naturlig å lage en regresjonsmodell, der søkertall (generelt, førstevalg og kvalifiserte) er den avhengige variabelen, og de forskjellige karaktersnittene er uavhengige variabler. En regresjonsmodell er en lineær modell som har en avhengig variabel som uttrykkes ved bruk av en eller flere uavhengige variabler (Freund et al., 2006). Modellen tar en form lik likning (9), der β_0 er intercept, det vil si et gjennomsnitt hvis alle uavhengige variabler (x) er lik null. $x_1$og x_2 er de uavhengige variablene, og ϵ er en såkalt error term, som vil si at den eksisterer for å ta hensyn til forstyrrelser som ikke tas opp av modellen.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad (9)$$

Til å begynne med ble det satt sammen en lineær modell med førstevalgs søkere som avhenig variabel. Etter at modellen er satt sammen testes de uavhengige variablene for kollinearitet, for å finne ut om modellen burde endres. Følgelig, når noen av de uavhengige variablene vises å ha høy kollinearitet endres modellen for å ta hensyn til dette. Resultat på kollinearitetstesten vil presenteres og diskuteres senere i oppgaven. Kollinearitet kan forekomme av tre årsaker (Hill & Adkins, 2001).

Den første årsaken er at en uavhengig variabel innehar liten variasjon. Dette er et problem, siden hensikten med regresjonen er å måle effekten av endringer i den uavhengige variabelen på den avhengige variabelen. Hvis den uavhengige variabelen ikke varierer gjennom datasettet er det vanskelig å måle noen som helst effekt.

Den andre årsaken er at to uavhengige variabler kan ha en stor korrelasjon seg imellom. Problemet oppstår når man prøver å se på individuelle effekter fra variablene. De individuelle effektene er vanskelige å måle, siden endring i en av dem også påvirker den andre.

Den tredje årsaken er at to uavhengige variabler er lineært avhengige av hverandre. Differansen mellom denne årsaken og den andre årsaken er om de uavhengige variablene beveger seg i samsvar med hverandre (tredje årsaken) eller om bevegelse i den ene fører til en bevegelse i den andre (andre årsaken).

I modellen er søkere førstevalg avhengig variabel. Det oppstår sterk kollinearitet i modellen, og kollinearitetstesten tilsier at søkere (til IT studier totalt) er årsaken til dette, på grunn av en for høy VIF verdi. VIF verdien er et mål for kollinearitet, og modellen burde endres når VIF

verdiene er over 10. VIF verdien til søkere var på 26.8. Søkere ble inkludert for at modellen skulle ta hensyn til alle faktorer, men ble fjernet for å beholde et tema blant de uavhengige variablene, altså at alle uavhengige variabler er gjennomsnitts karakterer. Ut ifra dette ble det laget tre regresjonsmodeller, en med søkere til IT-studier som avhengig variabel, en med kvalifiserte søkere som avhengig variabel, og en med søkere førstevalg som avhengig variabel.

Hvilken av de tre nevnte årsakene var grunnen til den sterke kollineariteten?

Rent intuitivt kan det ikke være den som oppstår som følge av for liten variasjon, om enn for den årsaken at det er ganske stor endring i søkere over tid. Vi ser på søkere over tid siden det er den uavhengige variabelen med høyest VIF verdi. Det største registrerte antallet er tross alt mer enn dobbelt så stort som det minste registrerte antallet når det kommer til søkere (19 106 og 7130). Variasjonen er mindre når det kommer til infotek snitt (som det viste seg at var sterkt korrelert med søkere), der det største tallet kun er en økning på omtrent 12% fra det minste tallet (fra 3.925 til 4.407), men siden det ikke var problem med kollinearitet etter å ha modifisert regresjonsmodellen til å ikke inkludere søkere virker det som om dette ikke var relevant.

Imellom de to andre alternativene, stor korrelasjon eller lineær avhengighet, virker det som om korrelasjon er det store problemet. På grunn av de store differansene mellom de registrerte tallene er det lite sannsynlig at problemet oppstår på grunn av lineær avhengighet. Årsak er derved trolig korrelasjon. En korrelasjonstest viser en korrelasjonsverdi på 0.946, som er veldig høyt. Av den årsak måtte en av variablene fjernes.

Det neste steget blir å sjekke korrelasjon mellom de resterende uavhengige variablene. Korrelasjonsverdiene for R2 og S2 er 0.89, R2 og Infotek er 0.85, og Infotek og S2 er 0.80. Korrelasjonene er ikke like stor som mellom søkere til IT-studier og infotek snitt, men de er fortsatt relativt store. Ettersom det ikke er kollinearitet vurderes det som akseptabelt. Siden alle de uavhengige variablene er gjennomsnittlige karakterer, gir det mening at det er høye korrelasjonsverdier dem imellom, særlig mellom R2 og S2, som begge er matematiske fag.

For å sjekke relevansen til de tre uavhengige variablene tas en lineær hypotese test. Til å begynne med testes hvorvidt noen av variablene kan være lik null, det vil si at de ikke har stor effekt på den avhengige variabelen. Den avhengige variabelen somt testes først er søkere førstevalg.

Testen har en nullhypotese om at en eller flere uavhengige variabler er lik null. For å avgjøre dette beregner man en F-verdi for hypotesetesten, som settes opp mot en kritisk F-verdi. Hvis beregnet F-verdi er større enn kritisk F-verdi forkastes null-hypotesen.

Videre testes hvilken av de tre uavhengige variablene som kan være lik null, om noen. Testen gir at den beregnede F-verdien til R2 snitt og infotek snitt er større enn kritisk F-verdi, mens F-verdien til S2 er mindre. S2 er derved ikke relevant for søkere førstevalg i IT-studier. Testen gjennomføres på søkere og kvalifiserte søkere, og kommer frem til samme konklusjon om at S2 ikke er relevant.

Den neste, og siste, testen er å sjekke om den nåværende modellen som ikke inkluderer S2 er bedre til å predikere søkertallene enn en som inkluderer S2 og de andre mulige variablene i datasettet. Dette gjøres ved å lage en restricted modell, som inneholder kun R2 snitt og infotek snitt. Det lages også en unrestricted modell som inkluderer R2 snitt og infotek snitt, men også flere andre uavhengige variabler. Testen går ut på å kalkulere en ny F-verdi som setter restricted og unrestricted modellene imot hverandre. Denne delen gjentas for de tre lineære modellene ved å endre avhengig variabel mellom søkere, søkere førstevalg og kvalifiserte søkere. F-verdien er her definert ved likning (10), der SSER er avvik i restricted modellen, SSEU er avvik i unrestricted modellen, og DF1 og DF2 er degrees of freedom fra unrestricted modellen.

$$F - verdi = \frac{\frac{SSER - SSEU}{DF1}}{\frac{SSEU}{DF2}} = \frac{SSER - SSEU}{DF1} * \frac{DF2}{SSEU} \quad (10)$$

I alle tre tilfellene kom det samme resultatet frem: R2 snitt og infotek snitt var de beste predikatorene i datasettet.

3.2.2 Vurdering av metoden

Metoden som er brukt i denne oppgaven er ikke blitt brukt i annen forskning, men sammenhengen mellom matematikk og IT/IKT er dokumentert. Det er også flere svakheter med metoden som er blitt brukt. Dette kapitlet vil gjennomgå noen av valgene som er tatt, og hva som kunne blitt gjort annerledes.

I oppgaven brukes karaktersnitt i R2 og S2 fra videregående år 3. Årsaken til å bruke R-matte og S-matte i oppgaven er basert på de spesielle kravene om R1, eller S1 og S2 som tidligere nevnt. Men i oppgaven blir R2 brukt i stedet for R1. Årsaken til dette er for at data skal

komme overens med økonomisk teori. Teorien om humankapital tilsier at et individ bygger opp kapital av utdanning. Det vil si at nivået av humankapital for et individ avgjøres av det siste nivået med utdanning de har.

Når oppgaven skal sjekke humankapital (målt i karaktersnitt) opp mot søkertall, er det viktig at nivået av humankapital blir representert på det tidspunktet et individ søker på studie. På søketidspunktet, som er etter videregående skole år 3, vil kunnskap innen R-matematikk være avgjort av karakterer ifra R2, ikke R1. Årsaken til denne påstanden er at nivå av humankapital vil være et år forskjøvet hvis man tar hensyn til R1-matematikk, da dette nivået oppnås et år før individet søker studie. Oppgaven forutsetter da at alle som tar R1-matematikk også tar R2-matematikk, som ikke nødvendigvis er sant. Svakheten bak dette valget derimot, er at oppgaven ikke ender opp med å teste de spesifikke kravene nøyaktig. Denne nøyaktigheten ble nedprioritert til fordel for å være mer presis i henhold til økonomisk teori.

Neste problem omhandler valget av å teste karaktersnitt. De spesielle kravene handler ikke om at man må ha en spesifikk karakter for å komme inn på studie, bare at man må ha tatt fagene i løpet av videregående skole. Dette fremkommer ikke ved å se på karaktersnitt. Et alternativ ville vært å bruke antallet elever som tok de respektive fagene på videregående. Fordelen med dette er at man kunne fokusert eksklusivt på fagene som etterspørres til studie, som er R1 og kombinasjonen S1 og S2. Da ville det kun vært nødvendig å se på antall som tok S2 matematikk, og en tidsforskyvning med ett år av de som tok R1 matematikk (ettersom man tar R1 året før man søker). Dette ville vært en bedre måte å teste relevansen av de spesifikke kravene til universitetene, men ville ikke være like grunnet i økonomisk teori om humankapital og skill.

Tidligere forskning har brukt gjennomsnittlig karakter som mål (Bjorvatn et al., 2018), men Bjorvatn et al. brukte gjennomsnittlig karakter for å teste effekten av et karakter-krav i lærer-studie. I deres metode er bruken av gjennomsnittlig karakter et åpenbart valg. Å måle humankapital er generelt vanskelig, og riktignok er gjennomsnittlig karakter i fag en simplifisert måte å gjøre det på, men det er vanskelig å finne andre metoder. Sannsynligvis må det simplifiseres til en viss grad hvis man skal måle kompetanse.

I artikkelen om frafall i løpet av IKT-studie i Estland (Kori et al., 2015) ble matematikk score brukt for å måle sannsynligheten for at noen droppet ut av studie. Kori et al. brukte mattescore fra studieløpet derimot, mens denne oppgaven bruker karaktersnitt i matematikk fra før studie.

Det gir åpenbart ingen mening at karakterer i løpet av studie skal ha noen direkte effekt på søkertall, som forekommer før studie. De to metodene bruker like mål, men ikke i samme sammenheng.

I artikkelen om IKT-mangel i Sør Afrika (de Villiers et al., 2012) nevnes det derimot at matematikk før studie spiller en stor rolle. Ifølge (Makapela, 2007), som henvist til i de Villiers et al., er matematikk viktig for å prestere i IKT-relaterte studier.

Det kan være andre problemer relatert til de uavhengige variablene i modellene. Som nevnt tidligere ville det vært optimalt med flere målinger. Det setter eventuelle resultater under tvil at det er få målinger, som gjør det vanskelig å definitivt trekke en konklusjon ifra resultatene. Det at den endelige modellen også bare har to uavhengige variabler kan bidra til at modellen tar hensyn til for få faktorer, eller at de faktorene i modellen blir forstørret. Valget å inkludere karaktersnitt ifra informasjonsteknologi ifra videregående skole kan også virke rart, med tanke på at målet var å teste relevansen av matematikk karakterer. Årsaken til å inkludere denne variabelen er forklart tidligere, ved at det virker logisk at hvis karakter spiller en rolle må denne variabelen spille en rolle, men den kan også ha skapt unødig støy når oppgaven testet matematikk karakterene.

Det er også mulig at modellen burde tatt en annen form. Modellen som er brukt er en lineær modell, men det ville også vært mulig å bruke for eksempel en log-lineær modell. Det er mulig at en slik modell ville passet bedre med data som er brukt i oppgaven.

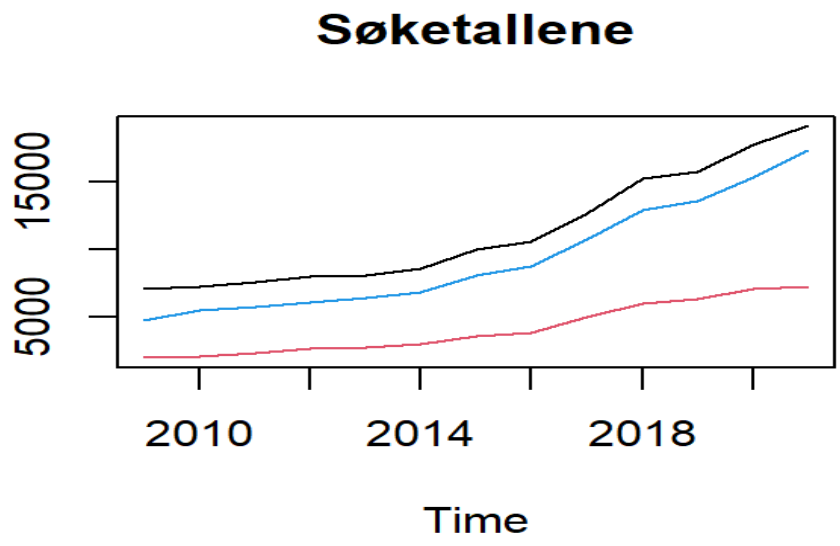
Generelt er det flere svakheter i metoden. Modellen kunne tatt en annen form, med andre variabler. Valget av variabler ender opp med å være en kombinasjon av spesifikke karakterkrav til studie, økonomisk teori, og intuisjon. En konsekvens av dette er at variablene som blir brukt i modellen ikke nødvendigvis er relevante, da for eksempel teori og spesifikke karakterkrav kan lede til at modellene tar forskjellig form. Som i oppgaven, teori og intuisjon tilsier at humankapital på søketidspunkt burde brukes som mål, mens det spesifikke kravet om R1 ikke nødvendigvis reflekteres i individet sitt nivå av humankapital på søketidspunkt.

Det neste kapitlet vil gå gjennom resultatene fra metoden i større detalj.

4 Resultater

Dette avsnittet vil oppsummere de forskjellige resultatene fra metode-delen av oppgaven. De relevante resultatene er de lineære modellene, kollinearitetstesten og korrelasjonstesten, resultater på den lineære hypotesetesten, og til slutt en test for å sjekke hvor god modellene er til å predikere de avhengige variablene (restricted og unrestricted modeller).

Visuell representasjon av søkertallene i figur 1:



Figur 1: grafisk presentasjon av søkertall

I figur 1 er sort søkere, blå er kvalifiserte søkere og rød er søkere førstevalg.

Til å begynne med, resultatene fra den lineære modellen for søkere førstevalg før kollinearitetstesten i tabell 1:

Tabell 1: lineær modell førstevalg m/ søkere og årstall

Søkere førstevalg lineær modell m/søkere og årstall

	Estimat	P-verdi
Intercept	-244400,00	0,00570
R2 snitt	1709,00	0,02343
S2 snitt	231,60	0,69781
Infotek snitt	-205,10	0,75963
INFOTEKN_søkere	0,29	0,000039
Årstall	118,30	0,00732

Multiple R-squared og Adjusted R-squared ekstremt nærme 1, der multiple R-squared er 0.9988 og adjusted R-squared er 0.9979. Det vil si at den lineære modellen kan forklare over 99% av variansen i den avhengige variabelen. Dette er veldig høyt så vi tar en kollinearitetstest.

Resultatene ifra kollinearitetstesten i tabell 2:

Tabell 2: kollinearitet førstevalgs søkere m/ søkere og årstall

Kollinearitet søkere førstevalg

	Tolerance	VIF
R2 snitt	0,0595	16,7970
S2 snitt	0,1194	8,3712
Infotek snitt	0,0601	16,6118
INFOTEKN_søkere	0,0373	26,8141
Årstall	0,4379	22,8365

Det ideelle er at VIF skal være lavere enn 10, og tolerance høyere enn 0.1. Som vi kan se i en korrelasjonstest er korrelasjonen mellom info_vgs (infotek snitt) og INFOTEKN_søkere (søkere) 0.946, og det avgjøres at en av de to variablene må fjernes. Etter kollinearitetstesten ble det avgjort å bare fokusere på snitt som uavhengige variabler. Det er verdt å nevne at modellen mister søkere og årstall, som begge er signifikante, men i senere modeller skal søkere brukes som avhengig variabel, så det blir fort rotete å ha søkere som uavhengig variabel i noen modeller, og avhengig variabel i andre modeller. Heretter gjelder bruken av bare snitt i modellene til søkere og kvalifiserte, der det ikke brukes årstall eller noen av de andre søkertallene som uavhengige variabler. Den nye kollinearitetstesten gir VIF lavere enn 10 og tolerance høyere enn 0.1. Kollineariteten mellom variablene er identiske for de tre modellene, og ser nå ut som i tabell 3:

Tabell 3: kollinearitet søkere, førstevalgs søkere, kvalifiserte søkere

Ny kollinearitet

	Tolerance	VIF
R2 snitt	1,1501	6,6617
S2 snitt	0,1931	5,1775
Infotek snitt	0,2630	3,8021

Heretter kommer de lineære modellene for de forskjellige søkertallene.

Lineær modell for søkere førstevalg er i tabell 4:

Tabell 4: lineær modell førstevalgs søkere

Søkere førstevalg lineær modell

	Estimat	P-verdi
Intercept	-42423,00	0,000000239
R2 snitt	6589,00	0,00163
S2 snitt	-1070,00	0,56500
Infotek snitt	5806,00	0,00108

Lineær modell før søkere er i tabell 5:

Tabell 5: lineær modell søkere

Søkere lineær modell

	Estimat	P-verdi
Intercept	-92295,12	0,00000221
R2 snitt	11340,93	0,02422
S2 snitt	55,23	0,99153
Infotek snitt	13979,05	0,00297

Lineær modell for kvalifiserte søkere er i tabell 6:

Tabell 6: lineær modell kvalifiserte søkere

Kvalifiserte søkere lineær modell

	Estimat	P-verdi
Intercept	-90417,00	0,00000264
R2 snitt	11383,00	0,0239
S2 snitt	-1837,00	0,7251
Infotek snitt	14685,00	0,0022

Kommentar: Intercept burde trolig ikke være negativ, siden antall søkere ikke kan være negativ. Dette kan være et tegn på svakheter i modellen. Vi ser også at S2 snitt er særdeles lite signifikant, spesielt i modellen som tar for seg alle søkere til IT-studier.

Lineær hypotesetest for å sjekke om de individuelle variablene er lik null er i tabell 7:

Tabell 7: lineær hypotesetest førstevalgs søkere, søkere og kvalifiserte søkere

Lineær hypotesetest

	R2 = 0	S2 = 0	Info_vgs = 0
Søkere førstevalg	19,7040	0,3568	22,3600
Søkere	7,3130	0,0001	16,2410
Kvalifiserte søkere	7,3624	0,1316	17,9090

Kritisk F-verdi er 3.86. Hvis kalkulert F er større enn kritisk F forkaster vi nullhypotesen om at variablene som testes er lik null. I de tilfellene der Kalkulert F-verdi er lavere enn kritisk verdi, betyr det at vi beholder nullhypotesen om at variabelen er lik null. I alle modellene gjelder dette for variabelen S2_snitt.

Resultat fra restricted og unrestricted modell testene er i tabell 8:

Tabell 8: Restricted og unrestricted modeller førstevalgs søkere, søkere og kvalifiserte søkere

Restricted og unrestricted

	Kritisk F-verdi	Kalkulert F-verdi
Søkere førstevalg	3,971	1,035
Søkere	3,971	0,307
Kvalifiserte søkere	3,971	0,254

Resultatet tilsier at restricted modellene predikerer bedre enn unrestricted modellene. Ut ifra dette avgjøres det at R2 snitt og infotek snitt er de eneste to viktige uavhengige variablene for modellene.

Nye lineære modeller:

Søkere førstevalg er i tabell 9:

Tabell 9: ny lineær modell førstevalgs søkere

Søkere førstevalg endelig lineær modell

	Estimat	P-verdi
Intercept	-43402,00	0,0000000095
R2 snitt	5995,00	0,000221
Infotek snitt	5670,00	0,000664

Søkere er i tabell 10:

Tabell 10: ny lineær modell søkere

Søkere endelig lineær modell

	Estimat	P-verdi
Intercept	-92245,00	0,000000121
R2 snitt	11372,00	0,003220
Infotek snitt	13986,00	0,001500

Kvalifiserte søkere er i tabell 11:

Tabell 11: ny lineær modell kvalifiserte søkere

Kvalifiserte søkere endelig lineær modell

	Estimat	P-verdi
Intercept	-92098,00	0,000000132
R2 snitt	10364,00	0,005910
Infotek snitt	14453,00	0,001270

Multiple R-squared og adjusted R-squared i alle modellene er tilnærmet uendret som i tabell 12:

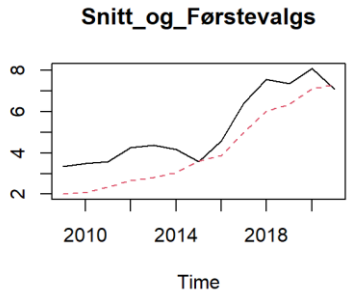
Tabell 12: multiple R-squared og adjusted R-squared, gammel og ny lineær modell, førstevalgs søkere, søkere og kvalifiserte søkere

	Modeller som inkluderer S2		Modeller uten S2	
	Multiple R	Adjusted R	Multiple R	Adjusted R
Søkere førstevalg	0,9750	0,9666	0,9740	0,9687
Søkere	0,9579	0,9438	0,9579	0,9495
Kvalifiserte søkere	0,9559	0,9412	0,9553	0,9463

Dette er et klart tegn på at det å fjerne S2_snitt som uavhengig variabel var riktig avgjørelse.

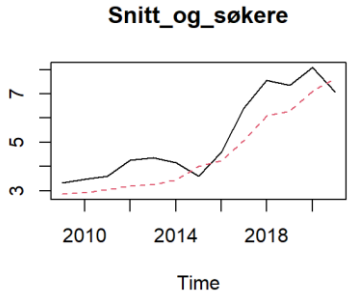
Vi finner at av de to mattefagene er R2_snitt signifikant. Variabelen har høyest signifikansnivå for søkere førstevalg, og litt lavere signifikansnivå for søkere og kvalifiserte søkere.

Visuelle sammenlikning av bevegelse i R2_snitt over tid og bevegelse i søkertall over tid:



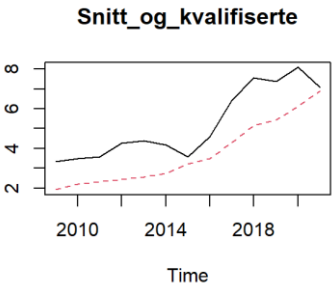
Figur 2: R2 snitt over tid og førstegangs søkere over tid

Figur 2 viser sammenhengen mellom R2 snitt og førstevalgs søkertall.



Figur 3: R2 snitt over tid og søkere over tid

Figur 3 viser sammenhengen mellom R2 snitt og søkere.



Figur 4: R2 snitt over tid og kvalifiserte søkere over tid

Figur 4 viser sammenhengen mellom R2 snitt og kvalifiserte søkere.

Siden oppgaven satte søkelys på matematikk, er det ikke viktig å vise infotek snitt opp mot søkertallene. Søkertallene er blitt nedjustert for å tydeligere visualisere bevegelse i de to

variablene. Siden tallene i y-aksen ikke er viktige spiller det ingen rolle at figur 3 har andre mål.

5 Diskusjon

Hvordan skal man tolke resultatene? En interessant ting å notere seg er at man kan se ifra figur 1 at søkere og kvalifiserte søkere beveger seg nesten identisk. Det vil si, det er omtrent en like stor andel av søkere som er kvalifisert, uavhengig av hvor mange som faktisk søker. Det kan forklare at resultatene, som blir diskutert senere i dette kapitlet, er veldig like for søkere og kvalifiserte søkere.

Den første modellen er en lineær modell av søkere førstevalg, som avhenger av snittkarakterer, søkere og årstall. Planen var å ta hensyn til det som virker å være en konstant stigning i søkere førstevalg over tid, ved å modellere at søkere førstevalg er en gitt andel av søkerne. Men, siden søkere (totalt for IT-studie) forklarer en så stor del av søkere førstevalg fører det til at det er vanskelig å faktisk måle viktigheten av de andre variablene, som er snittkarakterene. Det at søkere var så signifikant ga derfor veldig mye mening, men siden planen var å skape en lik modell for både kvalifiserte søkere, og søkere, var det vanskelig å beholde denne variabelen i modellen. Valget om å fjerne årstall falt da for å bare beholde snittkarakterer, slik at modellene er tematisk konsekvente. Som vi også kan se, så var korrelasjonen veldig høy mellom søkere og infotek snitt, og det ga mer mening å beholde infotek snitt i modellen enn søkere.

I tillegg, når kollinearitetstesten ble gjennomført var søkere det alternativet med høyest VIF nivå. Etter at søkere og årstall ble fjernet fra modellen var VIF nivået for de resterende variablene akseptable.

Når det kommer til de tre lineære modellene er trenden likedan, men til varierende grad. I samtlige av modellene er intercept veldig lav, og veldig signifikant. Infotek snitt og R2 snitt er signifikante til en lavere grad, mens i samtlige er S2 snitt ikke signifikant. Unikt for modellen til søkere (tabell 5) er hvor lavt estimatet til S2 snitt er sammenliknet med R2 snitt og infotek snitt. S2 er på 55, mens de andre to begge er høyere enn 11 000. Dette reflekteres i en p-verdi på 0.99, som tilsier at S2 snitt variabelen er 99% sannsynlig å være lik null i modellen. Det faktum at intercept er negativ i samtlige av modellene er litt bekymringsverdig, siden intercept er det punktet der grafen krysser y aksen. Ettersom antall søkere ikke kan være negativt, gir det liten mening at intercepten er negativ, spesielt gir det liten mening at det er veldig signifikant at den er så lav.

At S2 snitt ikke har en effekt på søkertallene kommer også frem ifra den lineære hypotesetesten. Ut ifra denne testen står R2 snitt mye sterkere i modellen for søkere førstevalg enn for de andre to modellene. Hvis man ser på R2 snitt som et mål for kompetanse ville man kanskje ha forventet at effekten var større på kvalifiserte søkere, slik som ble nevnt i hypotesen. Men årsaken til at det ikke er tilfellet kan være at det å være kvalifisert, per de spesielle studiekravene til universitetene, ikke påvirkes av evnen innen R2, men i stedet R1 og kombinasjonen av S1 og S2. Da burde man heller ha forventet å se en større effekt ifra S2 snitt, som er et av de spesifikke kravene. S2 var ikke signifikant i noen av modellene, og var etter testene best egnet til å predikere søkere, ikke kvalifiserte søkere. Årsak til at S2 snitt ikke er signifikant kan være at kravene som stilles er at et individ har hatt faget på videregående, men at det ikke stilles spesifikke karakter krav. Det kan også være at de som tar S2 matematikk på videregående skole bare generelt ikke søker til IT-studier.

Resultatet ifra testen der man setter opp en restricted modell og en unrestricted modell tilsier at R2 og infotek er beste måte å måle de forskjellige søkertall-modellene, i alle fall basert på de tilgjengelige variablene. Det utelukkes ikke at det kan være andre variabler som ikke er tatt med i beregningene, som predikerer søkertall like bra eller bedre. Dette resultatet samfaller med de tidligere testene som er gjennomført. Det er verdt å nevne at i disse testene brukes ikke søkere som uavhengig variabel (slik som i den aller første modellen i oppgaven), men i stedet brukes totalt antall søkere med alle som har søkt på et hvilket som helst studie som uavhengig variabel i modellen for søkere. Likeså brukes totalt antall kvalifiserte søkere, som er alle som har søkt på et studie de er kvalifisert til, som uavhengig variabel i modellen for kvalifiserte søkere. Og til slutt brukes totalt antall førstevalgs søkere som uavhengig variabel i modellen for førstevalgs søkere.

Når det så blir laget en ny lineær modell ender man opp med samme problem med inetercept som enda er negativ og veldig signifikant. Noe annet som man finner er at R2 snitt og infotek snitt har relativt like effekter på modellene. Men der infotek snitt er mer signifikant når det kommer til søkere og kvalifiserte søkere, er R2 mer signifikant når det kommer til søkere førstevalg. Differansene mellom de to variablene er ikke så store, så dette spiller ikke nødvendigvis en rolle. Men det kan virke rart, basert på det som er nevnt tidligere oppgaven med at det ikke er logisk at matematikk skal være mer signifikant i en modell om IT-studie enn infotek snitt.

I den nye lineære modellen kan vi også se at R-squared (multiple og adjusted) er relativt uendret, som tilsier at det var riktig å ikke inkludere S2 i modellen, siden variabelen hadde så liten effekt. Multiple R-squared går enten ned eller er uendret, som gir mening siden multiple bare tar hensyn til prediksjonseffekt, og alltid vil anta at en variable bidrar til en viss grad. Å fjerne en variabel vil da uansett føre til en lavere multiple R-squared. Adjusted R-squared går derimot opp (marginalt), siden den tar hensyn til antall variabler i tillegg. Siden S2 snitt ikke hadde så stor effekt, blir adjusted R-squared høyere. Dette skjer fordi modellen sin evne til å predikere avhengig variabel er tilnærmet uendret, men antall uavhengige variabler går ned.

Figur 2, 3 og 4 har alle fått søkertallene nedjustert med samme faktor, der linjen bare brukes til å visuelt vise sammenheng mellom R2 snitt og søkertall. Denne nedjusteringen er å dele på 1000. R2 snitt blir deretter korrigert for å få like verdier slik at begge kan vises ved siden av hverandre på en graf. Korrigeringen skjer ved å subtrahere snitt-nivået med 3.5, for så å multiplisere med 10. Det gjøres bare for å presentere snitt og søkertall ved siden av hverandre. Søkertall er naturlig større enn søkere førstevalg og kvalifiserte søkere. En konsekvens er at tallene i y-aksen blir annerledes, men de er uviktige. Et problem er at det visuelt kan se ut som om R2 passer bedre til en av søkere og kvalifiserte søkere, selv om bevegelsene i de to variablene er tilnærmet identiske. Det ser vi på figur 1 der forholdet mellom søkere og kvalifiserte søkere er konstant likt. Visuelt ser R2 veldig viktig ut for førstevalgs søkere, der foruten to avvikende observasjoner (2015 og 2021) ser det ut som bevegelsene i grafen er svært like.

Det er vanskelig å komme frem til en konklusjon ut ifra resultatene. Det at det er veldig få målinger gjør at det er noen tvil rundt om resultatene kan repliseres over tid, og en eventuell konklusjon vil innebære en viss usikkerhet.

6 Konklusjon

Basert på resultatene kan vi konkludere med at hvis man skal bruke matematikk karakter som mål på kompetanse ved søketid, og til å predikere søkertall innen IT-studier, så er det karakter i matematikk R2 som passer best. R2 snitt var best for å predikere søkere førstevalg med størst signifikans, og var litt mindre signifikant for søkere og kvalifiserte søkere. Dette passer ikke med forventningen om at matematikk ville være bedre å predikere kvalifiserte søkere. Litt overraskende er det at S2 snitt ikke var signifikant i noen av modellene, ettersom S2 er et av de spesielle kravene for IT-studier.

Data og metode er ikke feilfri. Spesielt med tanke på at det er få målinger er det vanskelig å konkludere med at resultatene reflekterer virkeligheten. Konklusjonen til spørsmålet «Kan karaktersnitt i matematiske fag brukes til å predikere antall søkere til IT-studier?» er derved at det er større sannsynlighet for at R2 snitt kan predikere antall søkere til IT-studier enn at S2 snitt kan predikere antall søkere til IT-studier, men det er ikke definitivt bekreftet.

Referanseliste

Bjorvatn, K., Huse, I., & Nilsen, E. O. (2018). Er matematikk viktig for å lykkes i lærerstudiet. *Samfunnsøkonomen*, 43.

Breiman, L. (1996). *Bias, variance, and arcing classifiers*.

Brigden, A., & Thomas, J. (2003). What does economic theory tell us about labour market tightness?

Cahuc, P., Carcillo, S., & Zylberberg, A. (2014). *Labor Economics* (Second edition ed.). The MIT Press.

Cappelli, P. (2005). Will there really be a labor shortage? *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management*, *44*(2), 143-149.

Cedefop. (2016). *Skill shortages in Europe: Which occupations are in demand - and why*. Cedefop. Retrieved 24.05.2022 from https://www.cedefop.europa.eu/en/news/skill-shortages-europe-which-occupations-are-demand-and-why

de Villiers, C., Johnson, R., & Cremer, P. (2012). South African ICT skills deficiency.

Docquier, F., & Rapoport, H. (2012). Globalization, brain drain, and development. *Journal of economic literature*, *50*(3), 681-730.

Duell, N. (2020). Skills shortages and labour migration in the field of information and communication technology in Canada, China, Germany and Singapore.

Eggen, F. W., Måøy, J., Røtnes, R., Norberg-Schultz, M., & Steen, J. I. (2021). *Norges behov for IKTkompetanse i dag og*

framover. S. Analyse. https://www.ikt-norge.no/wp-content/uploads/2021/01/r1-2021-behov-for-og-tilbud-av-ikt-kompetanse.pdf

Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis*. Elsevier.

Goldin, C. D. (2016). Human capital.

Hassouna, F., & Al-Sahili, K. (2020). Practical minimum sample size for road crash time-series prediction models. *Advances in civil engineering*, *2020*.

Hill, R. C., & Adkins, L. C. (2001). Collinearity. *A companion to theoretical econometrics*, 257-278.

Jones, C. I., & Vollrath, D. (2013). *Introduction to Economic Growth* (J. Repcheck, Ed. third ed.). Norton.

Kori, K., Pedaste, M., Tõnisson, E., Palts, T., Altin, H., Rantsus, R., Sell, R., Murtazin, K., & Rüütmann, T. (2015). First-year dropout in ICT studies. 2015 IEEE Global Engineering Education Conference (EDUCON),

López-Bassols, V. (2002). ICT skills and employment.

Mark, M. S., Tømte, C., Næss, T., & Røsdal, T. (2017). IKT-sikkerhetskompetanse i arbeidslivet–behov og tilbud.

McGrath, J. (2021). *Report on Labour*

Shortages and

Surpluses. E. L. Authority. https://www.ela.europa.eu/sites/default/files/2021-12/2021%20Labour%20shortages%20%20surpluses%20report.pdf

Naess, T. (2020). Master's degree graduates in Norway: Field of study and labour market outcomes. *Journal of Education and Work*, *33*(1), 1-18.

NTNU. (2022). *Informatikk - bachelorprogram 3-årig, Trondheim*. NTNU. Retrieved 06.06.22 from https://www.ntnu.no/studier/bit/opptak

Parviainen, P., Tihinen, M., Kääriäinen, J., & Teppola, S. (2017). Tackling the digitalization challenge: how to benefit from digitalization in practice. *International journal of information systems and project management*, *5*(1), 63-77.
Richardson, S. (2007). *What Is a Skill Shortage?* ERIC.
Ringard, Å., Sagan, A., Sperre Saunes, I., Lindahl, A. K., & Organization, W. H. (2013). Norway: health system review.
Taherdoost, H. (2017). Determining sample size; how to calculate survey sample size. *International Journal of Economics and Management Systems*, *2*.
UiT Norges Arktiske Universitet. (2022). *Informatikk - bachelor*. UiT Norges Arktiske Universitet. Retrieved 06.06.2022 from https://uit.no/utdanning/program/279505/informatikk_-_bachelor
Universitetet i Oslo. (2022). *Informatikk: programmering og systemarkitektur (bachelor)*. Universitetet i Oslo. Retrieved 06.06.2022 from https://www.uio.no/studier/program/informatikk-programmering/opptak/
Utdannings-direktoratet. (2021). *Karakterstatistikk for videregående skole*. https://www.udir.no/tall-og-forskning/statistikk/statistikk-videregaende-skole/karakterer-vgs/
Zuppo, C. M. (2012). Defining ICT in a boundaryless world: The development of a working hierarchy. *International journal of managing information technology*, *4*(3), 13.