



Universitetet
i Stavanger

HANDELSHØGSKOLEN VED UIS
BACHELOROPPGAVE

STUDIUM:

Bachelor i økonomi og administrasjon

OPPGAVEN ER SKREVET INNEN FØLGENDE TEMATISKE
RETNING:

Samfunnsøkonomi

TITTEL:

Hvordan vil kriminalitet aggregert mellom 2002-2020 påvirke lønn i 2021 i Norge?

ENGELSK TITTEL:

How will crime aggregated between 2002-2020 impact wages in 2021 in Norway?

FORFATTERE:

Kandidatnr:

1523

.....

1555

.....

.....

Navn:

Roem Daniell Malik Skjæveland

.....

Tobias Waage Thengs

.....

.....

VEILEDER:

Sara Helene Rønningstad

Forord

Denne bacheloroppgaven står som den avsluttende oppgaven før vi går av Handelshøyskolen ved Universitetet i Stavanger med vår bachelor innenfor økonomi og administrasjon.

Først vil vi takke vår veileder, Sara Helene Rønningstad, for å ha guidet oss gjennom oppgaven. Hun har hjulpet oss enormt gjennom veiledning og har vært tilgjengelig mer eller mindre til all tid. Sara Helenes forståelse og engasjement for emnet har bidratt til en kjekk og spennende skriveprosess.

Vi vil også takke familie, venner og medstudenter for all hjelp gjennom oppgaven. Fra hjelp til å lese over tekst, til nyttige innspill og vinklinger eller forståelse for temaer, men ikke minst deres støtte gjennom oppgaven og studiet som en helhet.

Det rettes også en takk til forelesere ved universitetet som har gitt oss en solid akademisk plattform gjennom bachelorgraden. Deres videreformidlet kunnskap og entusiasme har gitt oss byggesteinene til å realisere denne oppgaven.

Sammendrag

Kriminalitet kan utgjør en betydelig utfordring for samfunnet, med omfattende økonomiske konsekvenser. Denne studien undersøker hvordan kriminalitet påvirker lønn i Norge ved å analysere data fra perioden 2002 til 2020 og sammenligne det med lønnsdata for 2021. Fokuset havner på å forstå de økonomiske konsekvensene av kriminalitet i ettertid, med mål om å belyse utfordringer og behov i samfunnet. Vi fant at kriminalitet har en signifikant negativ innvirkning på lønn, hvor den ikke var veldig stor. Teorier som humankapital og signaleffekten blir tatt i bruk til å forklare hvorfor det var en såpass liten effekt av kriminalitet på lønn i Norge.

Abstract

Crime can pose a significant challenge to society, with extensive economic consequences. This study investigates how crime affects wage in Norway by analyzing data from the period 2002 to 2020 and comparing it with wage data for 2021. The focus is on understanding the economic consequences of crime in the aftermath, aiming to shed light on societal challenges and needs. We found that crime has a significant negative impact on wages, although it was not very large. Theories such as human capital and signaling effects are used to explain why there was such a small effect of crime on wages in Norway.

Innholdsfortegnelse

1 Introduksjon	7
2 Teori	8
2.1 Humankapitalteorien	8
2.2 Signaleffekten - Signalteorien	9
3 Hypotesene	9
4 Metode	11
4.1 Kvantitativ data	11
4.2 Deskriptiv statistikk	11
4.2.1 Utvalg	12
4.2.2 Ekstremverdi, Histogram & fordeling	12
4.2.3 Varians og standardavvik	13
4.3 Regresjon - OLS	13
4.3.1 Regresjon & regresjonskoeffisient	13
4.4 Antakelser ved regresjon	15
4.4.1 Linearitetsantakelsen	15
4.4.2 Homoskedastisitet til residualene	16
4.4.3 Multikollinearitet – uavhengighet	16
4.4.4 Normalfordeling	16
4.4.5 Antakelsene i microdata	17
4.5 Regresjonens t-verdi og p-verdi	17
4.6 Konvertering av logaritmiske koeffisienter	18
4.7 Hypotesetesting	18
5 Datamateriale	19
5.1 Behandlingssystemet	19
5.2 Datasettet	19
5.2.1 Variablers egenskaper	20
5.2.2 Variablene og definisjoner	21
5.2.3 Avgrensninger	23
5.2.4 Deskriptive resultater	24
5.2.5 Styrker og svakheter	26
6 Analyse	26
6.1 Regresjonsmodellene	26
6.2 Analyse av regresjonens antakelser	27
6.2.1 Linearitet	28

6.2.2 Heteroskedastisitet	28
6.2.3 Multikollinearitet	29
6.2.4 Normalitet	30
6.2.5 Konsekvenser ved brudd på normalitet & homoskedastisitet	32
6.3 Regresjonsanalyse	32
6.3.1 Kriminalitet.....	33
6.3.2 Kontrollert for alder	33
6.3.3 Kontrollert for kjønn	34
6.3.4 Kontrollert for utdanning.....	34
6.4 Hypoteseanalyse	35
7 Diskusjon.....	36
7.1 Alvorlighetsgrad?	36
7.1.1 Alvorligere straffereaksjoner	37
7.1.2 Tapt arbeidserfaring	37
7.2 Videre forskning.....	38
8 Konklusjon.....	40
9 Litteraturliste.....	41
10 Microdata kode (Vedlegg)	43

Innholdsregister

Tabell | Diagram | Histogram

TABELL 1 TESTER FOR ANTAKELSENE SOM GJENNOMFØRES I MICRODATA.....	17
TABELL 2 VARIABELOVERSIKT	21
TABELL 3 VARIABEL UTDANNING FRA DATASETET	22
DIAGRAM 1 KRIMINALITET FORDELT PÅ KJØNN.....	23
TABELL 4 DESKRIPTIV STATISTISK OVERSIKT FRA DATASETET	24
TABELL 5 HETEROSKEDASTISITET TESTER PÅ REGRESJONSMODELLENE	28
TABELL 6 VIF-TEST PÅ REGRESJONSMODELLER	29
HISTOGRAM 1 RESIDUALFORDELING FOR MODELL 1.....	30
HISTOGRAM 2 RESIDUALFORDELING FOR MODELL 2.....	30
HISTOGRAM 3 RESIDUALFORDELING FOR MODELL 3.....	30
HISTOGRAM 4 RESIDUALFORDELING FOR MODELL 4.....	31
TABELL 7 NORMALITET TESTER PÅ RESIDUALFORDELINGEN	31
TABELL 8 REGRESJONSMODELLENES RESULTATER	33
DIAGRAM 2 OVERSIKT OVER STRAFFEREAKSJONER I 2020	36
DIAGRAM 3 KRIMINALITET FORDELT PÅ UTDANNINGSNIVÅ.....	39

1 Introduksjon

Det er betydelige kostnader tilknyttet kriminalitet i Norge, hvor det var estimert til å koste samfunnet 144 milliarder kroner i 2019 (Eggen et al., 2022, s. 94-95). Utover det som er de overordnede samfunnskostnadene kan individuelle kostnader oppstå for de påførte ofrene, men også kriminalitetsutøveren. En tematikk som ofte undersøkes av regjeringen og forskningsinstitusjoner er påvirkningen kriminalitet har på samfunnet rundt dem og ofrene av kriminaliteten.

Kriminalitet, definert som “handlinger som bryter med gjeldende lovgivning” (Wikipedia, 2024, Kriminalitet), kan være drevet av ulike faktorer. Noen forskere peker på at tap av arbeid kan øke risikoen for kriminell adferd (Rege et al., 2009), mens andre har teorier om at økonomiske motivasjoner kan føre til vinningsbasert kriminalitet (Becker, 1968). En gjennomgående sammenheng i flere av disse forklaringene er økonomiens rolle, hvor det enten er tap av inntekt eller et ønske om økonomisk gevinst som motiverer til kriminelle handlinger.

En annen side av det økonomiske aspektet ved kriminalitet, som har en begrenset litteratursamling, er de økonomiske konsekvensene i etterkant av den avgjorte kriminaliteten. Dette er noe som kan være viktig siden kriminalitet kan medbringe en økonomisk konsekvens senere i livet som strekker lengre en straffereaksjonen på tidspunktet. En relevant publisering på dette emnet så på kriminalitets konsekvenser i alderen 15-24 år, hvor en av disse var lønnsmessige konsekvenser (Bhuller & Røgeberg, 2022, s. 68-69). Resultatene fra denne var at den gjennomsnittlige yrkesinntekten ble lavere for de straffede enn resten av befolkningen i aldersgruppen, men også at denne forskjellen økte med alderen.

Siden kriminalitetsraten blant unge mennesker er høy, hvor aldersgruppen 15-24 år stod for 29 prosent av alle siktelser i Norge i 2019 (Solberg, 2021, s. 21), mener vi dette er et viktig tema å utforske. Samtidig er det essensielt å ha et mer spredt fokus som inkluderer alle aldersgrupper. Dette skyldes at kriminalitet begått i ung alder ofte kan danne et mønster som fortsetter inn i voksenlivet, noe som understreker viktigheten av å forstå de økonomiske konsekvensene av kriminell atferd uavhengig av alder. Dette bidrar til å belyse hvordan kriminelle handlinger kan påvirke individers økonomi i ettertid.

Basert på det kunnskapshullet vi identifiserte rundt konsekvensene av kriminalitet, er målet med denne oppgaven å bidra til denne empiriske litteraturen. Vi skal derfor utforske

hvordan kriminalitet i Norge kan påvirke lønnen i 2021 etter at en person har blitt dømt for et lovbrudd mellom perioden 2002 til 2020, uavhengig av aldersnivået, utdanningsbakgrunn og kjønn. Oppgavens overordnede forskningsspørsmål er dermed, “Påvirker kriminalitet lønn i Norge?”. Blant funnene våres fant vi at kriminalitet har en signifikant effekt på lønn i Norge, selv om denne effekten ikke opptrer som veldig stor. Videre diskuterer vi også ulike faktorer som kan forklare hvorfor denne sammenhengen opptrer som den gjør, hvor vi finner at dynamikker i arbeidsmarkedet og systematiske forhold som politiattester er mulige årsaker.

2 Teori

Teori er avgjørende for diskusjonsdelen til oppgaven vår, hvor det setter et rammeverk til å diskutere sammenhenger basert på resultatene i oppgaven. Gjennom å inkludere eksisterende teoretiske perspektiver, vil vi ytterligere kunne drøfte hvordan lønn blir påvirket av kriminalitet i Norge. Dette kapitlet vil introdusere disse teoriene.

2.1 Humankapitalteorien

Teorien om humankapital er en av to teorier i oppgaven, som fokuserer på investeringer individer kan gjøre for å opparbeide større avkastning i form av lønn. Humankapitalteorien stammer opprinnelig fra en økonomisk teori som anser kunnskap, ferdigheter og erfaring som en fornybar ressurs som arbeidstakere kan investere i som kapital. Denne teorien betrakter disse ressursene som akkumulerbare, og antyder at investeringer i utdanning og kompetanseutvikling kan øke inntektsmulighetene til en person over tid (Becker, 1975).

Når vi introduserer humankapitalteorien ovenfor kriminalitetsselementet i oppgaven, er dette basert på en negativ anvendelse av humankapital. Personer som blir siktet for en kriminell handling kan møte ulike straffereaksjoner. En person som mottar en ubetinget dom, som medfører frihetsberøvelse, vil kunne oppleve tap av humankapital gjennom tapt arbeidserfaring og ansiennitet, som kan redusere jobbmulighetene.

Setter man i perspektiv tilbudene i Norske fengsler vil en ubetinget dom kunne føre til tap av jobbmuligheter, men ikke videreutdanning. Dette er fordi utdanningstilbudene i Norske fengsler skal være like som resten av samfunnet, hvor det tilbys yrkesfaglig eller annen høyere utdanning (Kriminalomsorgen, 2024). Som vil si at selv om kriminalitet kan

medføre utestengelse til å praktisere utdanningen, vil også muligheter for å opparbeide annen utdanning være positivt for humankapitalen.

2.2 Signaleffekten - Signalteorien

Den andre teorien som vi skal benytte oss av i oppgaven, er signaleffekten, en teori som vil sette søkelys på arbeidsmarkedet til forskjell fra humankapitalteorien som fokuserer på et mer individuelt nivå.

Signaleffekten er også en kjent økonomisk teori, som ble først introdusert i artikkelen “Job Market Signaling” av Micheal Spence i 1978. Artikkelen setter søkelys på hvordan individer i arbeidsmarkedet signaliserer sin produktivitet eller kvalifikasjoner til potensielle arbeidsgivere. Ettersom det er vanskelig å observere arbeidssøkeres sanne ferdigheter vil typiske signal som utdanning og arbeidserfaring være mer relevante indikatorer for arbeidsgiveres avgjørelse (Spence, 1978). Dette vil bety for personer med kriminell historikk, med en ubetinget straffereaksjon, tapt tid i arbeidsstyrken. Denne tapte tiden kan bidra til hull i CV-en og gjør det vanskeligere å signalisere sin pålitelighet, men også sine egenskaper til arbeidsgivere.

Videre vil yrker som krever politiattest for personer med en kriminell historikk også være vanskeligere å oppnå. Den kriminelle historikken kan indikere at personen har visse risikopreferanser eller egenskaper som arbeidsgivere ikke ønsker, men også ikke kan ansette på bakgrunn av lovverket.

3 Hypotesene

Hypotesene i oppgaven skal bidra på en systematisk måte å definere sammenhengers signifikans, noe som oppstår mellom de uavhengige variablene og den avhengige variabelen (Ubøe, 2017). Derfor skal vi i dette kapitlet introdusere hypotesene som skal besvares senere i analysen.

Hypotesene formuleres i strukturen H_0 , som er nullhypotesen og H_1 , som er alternativhypotesen. Hypotese 0 (H_0) er sann når man observerer ingen signifikant effekt på den avhengige variabelen, som gjør at vi aksepterer den hvis dette er indikasjonene (Ubøe, 2017, s. 181-186).

Vi vil benytte oss av to hypoteser, hovedvektene er utformet av forskningsspørsmålet, hvor disse hypotesene vektlegger kriminalitet.

HYPOTESE 1

H0: Kriminalitet har ingen signifikant sammenheng med lønn

H1: Kriminalitet har en signifikant sammenheng med lønn

Den første hypotesen undersøker om en endring i kriminell historikk har en sammenheng med lønn. Med denne hypotesen ønsker vi å undersøke effekten kriminalitet har på lønn i Norge, når ingen andre variabler er inkludert.

HYPOTESE 2

H0: Kriminalitet har ingen signifikant sammenheng med lønn når kontrollvariablene alder, kjønn og utdanning er inkludert

H1: Kriminalitet har en signifikant sammenheng med lønn når kontrollvariablene alder, kjønn og utdanning er inkludert

For den andre hypotesen har vi valgt å inkludere kontrollvariabler. Dette er gjort for å undersøke endringer som oppstår når vi eksempelvis legger til kjønn, og hva effekten da blir når en potensiell underliggende effekt av kjønn blir fjernet fra kriminalitetsvariabelen. Vi vet også at menn er overrepresentert i samfunnets lønnshierarki (Lekve & Arnesen, 2023), men også består i en stor andel av siktelsene avgjort i datasettet vårt (Kapittel 5). Dermed vil vi ved begge hypotesene få et klarere bilde av hvordan kriminalitet faktisk påvirker lønn i Norge, når typiske underliggende effekter som kjønn, alder og utdanning blir kontrollert for.

4 Metode

I økonomi og samfunnsforskning er statistikk et viktig verktøy for å adressere forskningsspørsmål. Derfor skal dette kapittelet fokusere på å henvise til de ulike statistiske metodene som er brukt i oppgaven, men også hvordan disse utgjør grunnlaget for å analysere dataen på en objektiv måte.

4.1 Kvantitativ data

Først og fremst er oppgaven basert på en kvantitativ tilnærming, noe som betyr at dataen vi bruker består av større mengder observasjoner, men med mindre beskrivende informasjon. Mer beskrivende informasjon vil man oppleve ved kvalitativ metode, som baserer seg på intervjuformer og mer forklarende tilnærminger.

Den kvantitative innsamlingsmetoden begrenser innsikten i eksempelvis personers individuelle psykisk helse, arbeidsmiljø eller andre faktorer som kan påvirke variabelen lønn (Johansen & Sundebye, 2019). Kvantitativ metode fokuserer på standardisering av observasjonene på bakgrunnen av mengden data. Dette gjøres ved deskriptiv statistikk som gjennomsnitt, varians eller median, uten nødvendigvis å forklare årsakene bak dem (Johansen & Sundebye, 2019).

4.2 Deskriptiv statistikk

Vi skal benytte dataanalyse, nærmere regresjon, derfor vil det være viktig å kunne oppsummere egenskaper ved datasettet som skal brukes. Dette gjør vi gjennom det vi kaller for deskriptiv statistikk, hvor store observasjoner blir standardisert for å gi en lettere forståelse av dataen som blir brukt. Denne type statistisk tilnærming søker å gi en objektiv besvarelse på endringer, variasjoner eller trender observert i dataen. De ulike metodene som benyttes i vår oppgave skal vi gå nærmere inn på i de neste kapitlene.

4.2.1 Utvalg

Representativitet er et viktig element i statistikk, det er sentralt at variablene i datasettet har observasjonsantall som på best mulig måte kan gi en representasjon av det vi kaller for populasjonen. Populasjonen er det totale antallet observasjoner du kan finne for en variabel uten avvik, og er vanskelig å måle. I fleste sammenhenger benytter man et utvalg fra populasjonen som ikke representerer totalen, men en andel. Denne andelen er ønskelig å være så stor som mulig slik at vi kan med større sikkerhet gjøre antakelser om populasjonen (Ubøe, 2017, s. 1-4).

4.2.2 Ekstremverdi, Histogram & fordeling

I observasjoner som har kontinuerlige verdier vil det kunne oppstå hva som kalles for ekstremverdier, dette er observerte verdier som oppstår sjeldent i datasettet.

Ekstremverdier kan tas hensyn til ved bruk av median. Medianen er midtpunktet i variabelens observasjoner når observasjonene er rangert i stigende kronologisk rekkefølge, og befinner seg i det som kalles for 50. percentilen eller 2. kvartil (Ubøe, 2017, s. 4-6). For å visualisere ekstremverdier, kan man anvende et histogram. Dette er en grafisk tilnærming som visualiserer alle observasjonene etter gruppering og frekvens (Ubøe, 2017, s. 7-8). Histogram vil også gi indikasjoner om fordelingen til observasjonene, men også hvilken observasjon som har den høyeste frekvensen, også kjent som modus eller mode på engelsk (Ubøe, 2017, s. 8).

Gjennom histogram vil vi kunne observere hva som er mest normalt, altså hvor observasjonene konsentrerer seg mest. Dette området blir betegnet som gjennomsnittet, og er verdien som representerer en likhetsfordeling mellom alle observasjonene (Ubøe, 2017, s. 9-10). Det er viktig å være oppmerksom på at ekstremverdier også kan trekke gjennomsnittet opp eller ned, som gjør at det kan være mindre realistisk at alle observasjonene i datasettet tar denne verdien og dermed et dårlig estimat (Eirik-Mathias et al., 2024).

4.2.3 Varians og standardavvik

For å kunne gi gjennomsnittet noen mer beskrivende karakteristikk kan man benytte varians og standardavvik. Standardavviket forklarer hvor mye den gjennomsnittlige observasjonen avviker fra gjennomsnittet og variansen tallfester volatilitet i dataen, altså et estimat på hvor mye opp og nedgangssvingninger som oppstår i dataen i gjennomsnitt (Ubøe, 2017, s. 10-11). Standardavviket foretrekkes ofte fordi det gir en bedre innsikt i avvikene vi observerer, ettersom enheten standardavviket uttrykkes i er lik dataen man bruker (Ubøe, 2017, s. 11-12).

4.3 Regresjon - OLS

Hovedvekten i oppgaven vil være på lineær regresjon. OLS (ordinary least squared) er en fellesbetegnelse for en lineær regresjon som oppstår i denne oppgaven (Ubøe, 2017, s. 249). Regresjon vil kunne gi indikasjoner til hvordan endringer i en persons kriminell historikk har en sammenheng med lønnen til personen, og vil bidra til å understøtte besvarelser rundt forskningsspørsmålet.

4.3.1 Regresjon & regresjonskoeffisient

I OLS er koeffisient en viktig betegnelse, som er en måling på samvariasjon mellom den uavhengige variabelen og den avhengige variabelen. Hvor den representerer endringen i variabelen Y (avhengige) hvis det oppstår en enhetsøkning i variabelen X (uavhengige) (Ubøe, 2017, s. 14-15).

ENKEL REGRESJONS MODELL

$$\hat{Y} = \beta_0 + X_i\beta_i + \epsilon$$

En lineær regresjon består av to variabeltyper. Hvor den første er den avhengige, som er den variabelen utgitt av modellen \hat{Y} , og er variabelen modellen søker å predikere. For å finne denne prediksjonen benyttes den andre variabeltypen, som er uavhengige variabler (X_i). Disse uavhengige variablene skal søke å gi en forklaring på variasjon i den avhengige variabelen (\hat{Y}) (Ubøe, 2017, s. 249-250). Forklaringsgraden er en måling på denne forklaringen, også kjent som R^2 . Denne måler i hvor stor grad de uavhengige variablene (X_i) forklarer variasjonen i den avhengige variabelen (Y) (Ubøe, 2017, s. 251-252).

Forklaringsgraden kan uttrykkes som forholdet mellom variansen til feilleddet (SSE) og den totale variansen til den avhengige variabelen (SST):

$$R^2 = 1 - \frac{SSE}{SST} \quad \beta \text{ SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad \beta \text{ SST} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Dette indikerer hvor mye av variasjonen i Y som kan forklares av variasjonen i X_i . Hvis R^2 er lavere enn 100%, indikerer det at en del av variasjonen i Y ikke blir forklart av de uavhengige variablene i modellen (Ubøe, 2017, s. 252). Det motsatte av forklaringsgraden er $1-R^2$, som representerer andelen av variansen i den avhengige variabelen som ikke blir forklart av de uavhengige variablene i modellen, men som egentlig eksisterer i populasjonen (Ubøe, 2017).

Utover forklarende sammenhenger kan vi i en regresjon også måle forskjellen mellom regresjonslinjen og de faktiske observasjonene, denne målingen kalles for residual ($R_i = Y_i - \hat{Y}_i$), eller feilleddet (Ubøe, 2017, s. 251-253). Hvor regresjonslinjen er en lineær linje utgitt av modellens prediksjoner, som viser sammenhengen mellom den uavhengige variabelen og den avhengige. Ved bruk av denne vil vi kunne si noe ytterligere om hvor mye feil modellen gjør i sine prediksjoner.

Feilleddet kan reduseres ved å legge til kontrollvariabler, dette er variabler som mistenkes å ha en sammenheng med avhengige variabelen. Kontrollvariabler kan også bidra til å fjerne underliggende effekter fra de uavhengige variablene, hvor eksempelvis sammenhengen mellom kriminalitet og lønn også kan være påvirket av kjønn. Dermed ved inkludering av kjønn blir variabelen mer uavhengig, og effekten av hver enkelt variabel blir mer presis (Ubøe, 2017).

MULTIPPEL REGRESJONS MODELL

$$\hat{Y} = \beta_0 + X_1\beta_1 + X_i\beta_i + \epsilon$$

Forskjellen mellom en enkel og multippel regresjon er antallet uavhengige variabler, og beta-koeffisientene (β_i). I multippel regresjonen vil alle de uavhengige variablene påvirke hverandres individuelle beta-koeffisienter, dette er fordi man fjerner samvariasjon som oppstår mellom variablene for å sikre uavhengighet, og innvirkning på den avhengige variabelen (Y) blir dermed endret.

For en multipl regressjon, i likhet med enkel, vil beta-koeffisientene ogs  representere stigningstallet. Ettersom det er flere variabler i en multipl m  man ta en forutsetning om at alle de andre uavhengige variablene i modellen er konstante.

LOG-LINE ER REGRESJONS MODELL

$$\ln(\hat{Y}) = \beta_0 + X_1\beta_1 + X_i\beta_i + \epsilon$$

I modeller der den avhengige variabelen (Y) er logaritmisk transformert, kalles den for "log-line er" (Stock & Watson, 2015, s. 318). Tolkningen av koeffisientene (β) i en slik modell tilsier at " kningen i X med en enhet er assosiert med $(100 * \beta_1)\%$ endring i Y " (Stock & Watson, 2015, s. 319). Dette vil si at hvis koeffisienten tar verdien 0,1, vil en enhets kning (1) i X medf re en 10% gjennomsnittlig  kning, men er relativ til den den logaritmiske fordelingen til Y .

4.4 Antakelser ved regresjon

N r man skal benytte regresjonsanalyse er det viktig at resultatene vi f r er riktige, og ikke lider av skjevheter. Ved   gjennomf re tester p  regresjonsmodellene vil vi sikre at prediksjonene vi f r har en h yere sannsynlighet for   v re riktige. Antakelsene som skal testes er Linearitet, Homoskedastisitet, Multikollinearitet og Normalitet (James et al., 2013, s. 92-102). Det er viktig   p peke at microdata har sine begrensninger i forhold til personvern, som eliminerer muligheter for   teste noen av antakelsene, noe som ogs  legger til grunn valget av disse fire. (Sikt, & SSB 2022).

4.4.1 Linearitetsantakelsen

Linearitet er en veldig sentral men intuitiv antakelse for en regresjonsmodell, denne referer til at man  nsker ett "rettlinjet forhold mellom prediktorene og responsen" (James et al., 2013, s. 92-93). I praksis ser vi ikke p  forholdet mellom disse to, men bruker heller residualene for   trekke konklusjoner.

I v r oppgave er det begrenset tilgang p  visualisering av Residual-plot (Scatter plot) p  bakgrunn av microdatas personvern. Denne kunne vist fordelingen av residualene i forhold til en line r linje utgitt av modellen. Siden vi ikke kan benytte denne, kan vi heller ikke kunne konkludere noe om dette i oppgaven.

4.4.2 Homoskedastisitet til residualene

Ved Homoskedastisitet menes det at residualene (R_i) i modellen bør ha en konstant varians og et gjennomsnitt som går mot verdien 0. Dette er fordi feilen modellen gjør i prediksjonene sine skal ha en lik spredning upåvirket av forskjellige verdier til de uavhengige (X_i) variablene, som vil si at modellen er like presis i sine prediksjoner uavhengig av hvilke verdier den avhengige (Y_i) variabelen tar. Er residualene sin varians ikke konstante opplever man heteroskedastisitet (James et al., 2013, s. 95-96).

4.4.3 Multikollinearitet – uavhengighet

For å sikre at variablene man bruker i regresjonen er av forklarende natur, er det også viktig at disse har en sann uavhengighet. Avviker man fra denne uavhengigheten opplever man multikollinearitet eller kollinearitet, forskjellen mellom disse er at multikollinearitet referer til når 3 eller flere uavhengige variabler i regresjonsmodellen avviker, mens kollinearitet referer til 2 eller flere (James et al., 2013, s. 99-101). For å oppdage om det eksisterer multikollinearitet kan man kjøre en såkalt VIF test, hvor denne utgir laveste verdi 1. “Som en hovedregel er VIF verdier over 5 og 10 indikasjon på problematisk mengde kollinearitet” (James et al., 2013, s. 101-102), noe som indikerer at multikollinearitet har en større sannsynlighet for å eksistere i datasettet.

4.4.4 Normalfordeling

Normalitet i regresjonsanalyse referer primært til antakelsen om en normalfordeling av residualene i regresjonsmodellen. Dette vil si at for enhver verdi de uavhengige variablene tar så bør residualene følge en form for normalfordeling. Det er da viktig å huske at residual er avviket mellom det vi observerer og det modellen predikerer, som vil si at hvis disse ikke er normalfordelt vil variasjonen mellom residualene være høy og prediksjonene modellen gir kan lide av skjevhet (Ubøe, 2017, s. 269-271).

4.4.5 Antakelsene i microdata

Når vi gjennomfører analysen i microdata, vil det være begrensninger på tester og visuelle oversikter. Derfor vil vi gi en kort introduksjon for hvilke statistiske tester som kan gjennomføres og hva disse testene forteller oss:

TABELL 1 TESTER FOR ANTAKELSENE SOM GJENNOMFØRES I MICRODATA

Type antakelse	Test	Resultatindikatorer
Linearitet	Ingen	Ingen metode tilgjengelig i microdata for å avkrefte eller bekrefte denne antakelsen
Homoskedastisitet	Breusch-Pagan Tester	H0 = Homoskedastisitet eksisterer og residualene har konstant varians H1 = Heteroskedastisitet eksisterer og residualene har ikke konstant varians Finner man P-verdi < 0,05 har man bevis for å avvise H0
Multikollinearitet	VIF Test	Tester hvor mye en enkelt uavhengig variabel blåses opp på bakgrunn av en korrelasjon mellom de andre uavhengige variablene. Ønsker at gjennomsnittlig VIF er nærmest 1 og kan da bekrefte at multikollinearitet ikke eksisterer
Normalitet	Residual histogram	Visuelt ønsker en normalfordeling, observasjoner konsentrert rundt gjennomsnittet.
Normalitet	Jarque-Bera Test (Normaltest)	H0 = Normalfordeling eksisterer i residualene H1 = Normalfordeling eksisterer ikke i residualene Finner man P-verdi > 0,05 har man ikke bevis for å avvise H0, og indikerer at variablene stammer fra normalfordeling.

Tabell 1 Normalitet antakelsen i histogram (Ubøe, 2017, s. 269-271), VIF test for multikollinearitet (James et al., 2013, s. 99-102), Breusch-pagan test på homoskedastisitet (Wooldridge, 2020, s. 268-279) Normalitet antakelsen fra jarque-bera test (Thadewald & Büning, 2004, s. 88-92).

4.5 Regresjonens t-verdi og p-verdi

I en regresjon vil man kunne finne svar på T-verdier og P-verdier. Hovedforskjellen mellom disse er at T-verdien er en statistisk test som tester forskjellen mellom observasjonene i variabelen og nullhypotesen (Ubøe, 2017, s. 203-205), og P-verdien er en betegnelse på hvor signifikant disse variablenes effekt er (Ubøe, 2017, s. 189-192).

Vanligvis er det forhåndsbestemte signifikansnivået satt til 5% (Ubøe, 2017, s. 177-179). Dette definerer området hvor p-verdier må falle under for å kunne avvise nullhypotesen. P-verdier representerer sannsynligheten for å observere en prediksjon fra modellen under nullhypotesen. Hvis P-verdien er større enn det forhåndsbestemte signifikansnivået

indikerer det at funnene ikke er signifikante, og vice versa hvis p-verdien er mindre en signifikansnivået (Ubøe, 2017, s. 177-179)

T-verdier beregnes ved å dividere variabelens koeffisient med standardavviket for å beregne en verdi som indikerer hvor stort den estimerte koeffisienten fra modellen avviker fra null, også kjent som nullhypotesen (Hartmann et al., 2023). Veldig høye t-verdier tyder ofte på veldig lave p-verdier i modellen, dette betyr også det at den tilsvarende prediktoren (X) har en veldig sterk effekt på den avhengige variabelen (Y).

4.6 Konvertering av logaritmiske koeffisienter

Vi nevnte tidligere for den log-lineære modellen at endringene tok utgangspunkt i den logaritmiske fordelingen til Y. I denne delen skal vi introdusere formelen for hvordan vi kan konvertere denne koeffisienten:

$$(e^{\beta_i} - 1) * 100\%$$

Denne verdien utgitt av formelen vil representere den prosentvise endringen i Y når den ikke er logaritmisk transformert, dette gjør at vi kan si noe gjennomsnittslønnen i Norge. Denne konverteringen er noe som vil bli brukt i analysedelen av oppgaven.

4.7 Hypotesetesting

Hypotesene som ble introdusert tidligere i oppgaven, er en fundamental del av oppgavens analyse, som tillater oss å konkludere om de observerte sammenhengene er signifikante på en objektiv måte. Etersom dette er en viktig del av oppgaven skal vi i dette kapittelet introdusere hva som kan påvirke gyldigheten av disse.

Vurderingen gjøres ved å evaluere sannsynligheten for å observere visse resultater under antakelsen av en nullhypotese (H_0). Disse resultatene kan føre til to typer feil, Type 1- eller Type 2-feil. (Ubøe, 2017, s. 177-179).

En type 1- feil oppstår når nullhypotesen (H_0) blir feilaktig forkastet. Dette indikerer en sammenheng som faktisk ikke eksisterer i populasjonen, men som våre resultater indikerer. Det er en feilaktig konklusjon som viser bevis for alternativ hypotesen (H_1), når vi egentlig ikke burde avvise nullhypotesen (Ubøe, 2017, s. 178). En type 2-feil oppstår når

nullhypotesen feilaktig beholdes, noe som skjer når en sammenheng som faktisk eksisterer i populasjonen ikke blir oppdaget (Ubøe, 2017, s. 177).

I vår oppgave har vi valgt å benytte et forhåndsbestemt signifikansnivå på 0,05, som betyr at vi aksepterer en feilmargin på 5%. Når p-verdien som oppnås i analysen er lavere enn dette signifikansnivået, forkaster vi nullhypotesen (Ubøe, 2017, s. 178-179).

5 Datamateriale

I dette kapittelet vil man beskrive og forklare hvordan datamaterialet våres er bygget opp og behandlet, det vil innebære alt fra variabelbruk til avgrensninger. Variablene vi bruker er uthentet av microdata's variabeldatabase (Sikt, & SSB, 2022, s. 9-15). Microdata har store begrensninger når det kommer til personvern på disse variablene, noe som gjør at visse avgrensninger og eventuelle innsikt i observasjonene er vanskelig (Sikt, & SSB, 2022, s. 69), (Sikt, & SSB, 2024).

5.1 Behandlingssystemet

Når vi utfører en statistisk analyse så ønsker man resultater som kan representere populasjonen best mulig (Stock & Watson, 2015). Derfor har vi i denne oppgaven valgt å benytte oss av microdatas database og behandlingstjeneste, som er utviklet i et samarbeid mellom Sikt og SSB (Microdata, 2024a). Denne tjenesten opererer som en portal for alle observasjoner innsamlet av SSB, som kan behandles i deres eget system underbygget av programvaren Python og Pandas (Sikt, & SSB, 2022, s. 5-6).

5.2 Datasettet

Nå som vi har introdusert hvor dataen vi skal bruke er behandlet og hva de eksterne begrensninger ved dette systemet er, vil vi nå gå mer i detalj på hvordan datasettet vårt er oppbygget. Kapittelet vil inkludere alt fra variabler til styrker og svakheter ved datasettet.

5.2.1 Variablers egenskaper

I databasen til Microdata finnes det ulike typer variabler, hver med sine individuelle egenskaper. De mest typiske egenskapene for våre variabler inkluderer perioden de gjelder for, variabeltype og temporalitet. Vi benytter oss av to hovedtyper variabler: numeriske og alfanumeriske. Numeriske variabler består kun av tall, enten heltall (integere) eller desimaltall. Alfnumeriske variabler inkluderer både tall og bokstaver som verdier (Merriam-Webster, 2024, alphanumeric).

Når det gjelder innsamlingsmetoden for variablene, bruker microdata definisjonen temporalitet, som kommer i fire grupperinger. Den første er forløp som inneholder sekvenser av hendelser. For eksempel, hvis en person med videregående utdanning velger å utdanne seg på høyskolenivå, vil denne endringen i utdanning oppdatere observasjonen i datasettet til høyeste utdanningsnivå registrert (Sikt, & SSB, 2022, s. 27-28).

Den neste typen temporalitet er fast, noe som betyr at det bare er én observasjon per enhet. For eksempel, for en variabel som kjønn, vil den ikke kunne telle observasjonen som både mann og kvinne, kun én av dem. Dermed kan hver observasjon bare oppta en plass i variabelen (Sikt, & SSB 2022, s. 27-28).

Tverrsnitt er den tredje typen, som har målinger på faste tidspunkter. Det vil si at det ikke er oppfølgende observasjoner over perioder. For eksempel, for kriminalitet vil dette representere at en avgjort siktelse i perioden ikke vil være med i statistikken neste år (Sikt, & SSB, 2022, s. 27-28).

Den siste typen er det som kalles akkumulert temporalitet. Dette er også en variabel som ikke strekker seg over år, men er en sum av alle innen grupperingen over en periode. Dette gjelder ofte for variabler som lønn, hvor den viser en akkumulert verdi av lønn for en gitt periode (Sikt, & SSB, 2022, s. 27-28).

5.2.2 Variablene og definisjoner

Alle variabler introdusert i dette delkapittelet er uthentet fra variabel databasen til microdata. Datasettet er koblet opp til databank no.ssb.fdb:26, som er registerdata fra SSB's statistikkproduksjon sist oppdatert 20.11.2023 (Sikt, & SSB, 2022, s. 26). Datasettet er produsert ved navnet lønn_krim og koden til dette er vedlagt oppgaven. Variablene er funnet ved hjelp av microdata's variabeloversikt som gir en deskriptiv forklaring tilhørende variabelen (Sikt, & SSB, 2022, s. 9-15). Oversikten over alle variabler brukt vises under:

TABELL 2 VARIABELOVERSIKT

NAVN	BRUKSOMRÅDE	TYPE	TEMPORALITET
KRIMINALITET_SIKTEDE_LOVANT	KRIMINALITET	Numerisk	Tverrsnitt
INNTEKT_WLONN	LØNN	Numerisk	Akkumulert
NUDB_BU	UTDANNING	Alfanumerisk	Forløp
BEFOLKNING_FOEDSELS_AAR_MND	ALDER	Numerisk	Fast
BEFOLKNING_KJOENN	KJØNN	Alfanumerisk	Fast

Tabell 2 Oversiktens informasjon er uthentet av microdata's variabeloversikt (Microdata, 2024b). Tabellen er utarbeidet i Excel.

Variablene er tilført datasettet ved funksjonen "import" (Sikt, & SSB, 2022, s. 16), hvor alle variabler er fra den samme databanken. Det tilhørende bruksområdet for variablene skal beskrives i de neste avsnittene.

Kriminalitet er basert på variabelen KRIMINALITET_SIKTEDE_LOVANT viser "antall avgjorte siktelser/lovbrudd til siktet person i løpet av året". Variabelen er tverrsnitt, og viser personer med "endelig avgjørelse i årsperioden han/hun er registrert og antatt som skyldig av påtalemyndighetene" (Microdata, 2024b). I datasettet vårt er variabelen definert som krim_sum, og inneholder denne variabelen summert, eller aggregert, fra perioden 31.12.2002 til 31.12.2020. Det er viktig å bemerke seg at variabelen inkluderer alle avgjorte siktelser, som tilsier at det kan være alt fra fartsbøter til økonomisk vinningskriminalitet. I oppgaven vil man referere til både kriminalitet og avgjort siktelse, men disse to begrepene omfavner hverandre med samme betydning.

Lønn, som fungerer som den avhengige variabelen i regresjonsanalysen, er hentet fra variabelbeskrivelsen med navnet INNTEKT_WLONN. Denne variabelen, med endelig definisjon ln_lønn i datasettet, representerer akkumulerte lønnsinntekter for kalenderåret,

og inkluderer “kontantlønn, skattepliktige naturalytelser, og syke- og fødselspenger” (Microdata, 2024b). Det er viktig å bemerke seg at lønn representerer en logaritmisk transformasjon av lønnsvariabelen, og importert med perioden 2021 til datasettet.

Utdanning utgjør den tredje variabelen, og er en kontrollvariabel med grunnleggende betegnelsen NUDB_BU. Dette er en forløpsvariabel som bruker NUS-koder fra 0 til 9, hvor 0 representerer ingen utdanning, og høyere NUS-koder indikerer høyere utdanningsnivå, med kode 9 som betegner uoppgitt utdanning i høyeste nivå. Det første tallet i utdanningsgrupperingen angir NUS-koden (Microdata, 2024b). I datasettet er denne variabelen navngitt gruppert_utdanning og inndelt i gruppene 0 til 6:

TABELL 3 VARIABEL UTDANNING FRA DATASETET

GRUPPE	RESPEKTIV UTDANNING	NUS GRUPPERING, FRA-TIL
0	Ukjent avvik	<099901
1	Ingen utdanning	99901
2	Forskole	099902 - 301110
3	Videregående	301111 - 611115
4	Bachelor	611116 - 701999
5	Master	711101 - 801999
6	PHD & annen høyeste utdanning	811101 - 999999

Tabell 3 NUS grupperingene er uthentet av microdata (Microdata, 2024b), respektive utdanningsgrupper er uthentet av microdata koden vedlagt oppgaven. Tabellen er utarbeidet i Excel.

Variabelen utdanning, refererer til dataene som er gyldige per avslutningen av perioden 31.12.2021, og omfatter alle personer hvor utdanningsnivået deres er registrert. Personer uten registrert utdanningsnivå blir derfor ikke inkludert i statistikken (Microdata, 2024b). Det ukjente avviket i gruppe 0 er det samme som gruppe 1, ingen utdanning i variabelen.

De siste variablene inkluderer alder og kjønn, som i likhet med utdanning er kontrollvariabler i regresjonsanalysen. Alder kom med definisjonen BEFOLKNING_FOEDSELS_AAR_MND og kjønn med definisjonen BEFOLKNING_KJOENN. Begge variablene er fast temporalitet. Alder viser "fødselsår og måned" og "omfatter bosatte i Norge" (Microdata, 2024b). Vi omkodet variabelen alder til alder_2021 i datasettet med utgangspunkt i personer fra 2021, og kjønnsvariabelen ble omkodet til kjønn.

For kjønnsvariabelen ble vi stoppet fra å bruke den originale kodeformen i regresjonen på grunn av alfanumeriske verdier, derfor ble denne omkodet til numerisk verdi ved “destring” kommando (Sikt & SSB, 2022, s. 48). Videre er kjønnsvariabelen originalt kodet med verdiene 1 for mann, og 2 for kvinne, og omfatter kun kjønn til “personer registrert bosatt i Norge” (Microdata, 2024b). Kjønn er en relevant kontrollvariabel ettersom fordelingen mellom menn og kvinner med minst en avgjort siktelse på seg er skjevfordelt, vist av diagrammet under er 79% av alle personer med en avgjort siktelse eller mer, menn.

DIAGRAM 1 KRIMINALITET FORDELT PÅ KJØNN

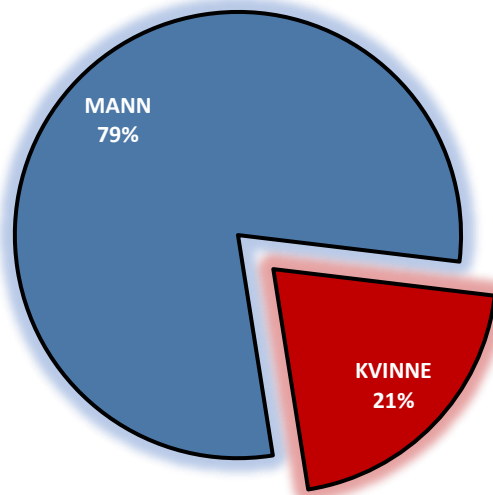


Diagram 1 Dataen er uthentet av microdata koden vedlagt oppgaven. Det viser fordelingen av menn og kvinner i datasettet som har avgjorde siktelse > 0. Diagrammet er utarbeidet i Excel.

5.2.3 Avgrensninger

Når vi arbeider med store variabler i datasettet er det behov for å gjøre visse avgrensninger, slikt at resultatene vi får i analysen kan besvare forskningsspørsmålet. Derfor skal vi i dette delkapittelet av datamaterialet introdusere avgrensningene gjort i datasettet vårt.

Den mest merkbare begrensning ble gjort på lønnsvariabelen (ln_lønn), hvor vi har avgrenset den til kun observasjoner med en lønn på 60 000 kr eller mer, samt en aldersbegrensning fra 15 år og oppover. Dette valget er begrunnet ved frikortgrensen for skatt i Norge for 2021 (Skatteetaten, 2024), samt den kriminelle lavalderen på 15 år for samme år (Lov om Straff, 2005, §20). Dette sikrer at observasjonene for lønn også blir best mulig representerer relativt til den forklarende variabelen kriminalitet, men samtidig

fjerner ekstremt lavlønnede observasjoner som ikke er representative for den gjennomsnittlige arbeidstakeren.

Videre har vi gjort en standardisert avgrensning i datasettet som stopper alle variabler på perioden 2021, men gjelder ikke kjønnsvariabelen, ettersom denne antas ikke å endres over tid. Denne periodeavgrensningen skyldes at den siste tilgjengelige kriminalitetsvariabelen for avgjorte siktelser (krim_sum) er i slutten av perioden 2020. Dermed ved å benytte 2021 får vi undersøkt hva den etterkommende effekten av de avgjorte siktelsene er på lønnsvariabelen. Tilsvarende er perioden begrenset for kontrollvariabelen alder, slik at alderen til observasjonene er relativ til året 2021 som referansepunkt og ikke deres nåværende alder.

5.2.4 Deskriptive resultater

Senere i oppgaven skal vi analysere resultatene, derfor vil vi vurdere datasettet på forhånd siden resultatene vil stamme fra dette. Vurderingen gjøres ved å påpeke styrker og svakheter, hvor vi benytter deskriptiv statistikk som gjennomsnitt, standardavvik, median og antallet observasjoner for å utvekle objektive indikasjoner til styrker eller svakheter

TABELL 4 DESKRIPTIV STATISTISK OVERSIKT FRA DATASETDET

Variabel	Gj.snitt	Std.avvik	Antall	0,01	0,25	Median (0,5)	0,75	0,99
alder_2021	42,153	13,622	270778 8	18	31	42	53	70
krim_sum	0,3413	1,1287	270778 8	0	0	0	0	8
gruppert_utdannin g	3,3481	1,1068	270778 8	0	3	3	4	6
ln_lønn	13,026 7	0,691	270778 8	11,1	12,7	13,2	13,5	14,4
lønn	553906	322800	270778 8	6710 0	33800 0	527000	69700 0	179000 0
kjønn	1,474	0,4993	270778 8	1	1	1	2	2

Tabell 4 Dataen er uthentet av resultater funnet ved microdata koden vedlagt oppgaven. Utgir deskriptive resultater ved datasettet. Tabellen er utarbeidet i Excel.

For alle variablene i tabellen (Tabell 4), ekskludert kriminalitet, observerer man at gjennomsnittet ligger relativt nært medianen, noe som kan indikere at det er mindre

påvirket av ekstremverdier. Oppsummerende fra tabellen fant vi følgende gjennomsnittsverdier, hvor alder var på 42år, lønn på om lag 553.000kr, utdanning på videregående (3) nivå, kjønn på 1,47 og kriminalitet med verdien 0,34. Variablen $\ln_lønn$ nevnes ikke fordi den er det samme som lønn, men logaritmisk transformert.

En merkbar variabel var kriminalitet, hvor denne hadde ett gjennomsnitt som er lavere enn verdien 1, altså en avgjort siktelse. Derfor kan vi, basert på dette gjennomsnittet, si at det er veldig lite personer som tar på seg en avgjort siktelse i datasettet, og dermed trekker ned gjennomsnittet til verdien observert. Det andre er kriminalitets avvik fra medianen på 0, sett i kombinasjon med øvre kvartilen (0,99) på 8, kan bety tilstedeværelse av ekstremverdier.

Videre for kontrollvariabelen alder ($alder_2021$) indikeres det en betydelig spredning, med et standardavvik på 13,6 år, noe som tyder på at representasjon av et bredt spekter av aldersnivå i datasettet. Dette bekreftes også av den øvre kvartilen (0,99) på 70 år, og nedre kvartil (0,01) på 18år. Den andre kontrollvariabelen, kjønn, kan også indikere en høy spredning relativt til gjennomsnittet, med et standardavvik på 0,4993. Men ettersom variabelen kun tar verdiene 1 og 2 er dette ikke uforventet. Denne observasjonen er dermed noe som tyder på at kjønn har en lik representativ fordeling i datasettet vårt.

Den siste kontrollvariabelen som er utdanning ($gruppert_utdanning$), viser et standardavvik på 1,1068 og et gjennomsnitt på om lag tregangeren dette. Dette betyr at gjennomsnittlig kan man forvente at den normale observasjonen kan både ha utdanningsnivå 2 og 4 også. Den øvre kvartilen (0,99) er også betydelig høyere enn gjennomsnittet, med en verdi på 6, mens den nedre kvartilen (0,01) er på 0. Dette kan videre indikere eksistens av ekstremverdier, altså få personer med veldig høy utdanning. Dette er ikke uforventet ettersom utdanning i høyere nivå er mer utfordrende og dermed mindre deltakere på slikt nivå.

5.2.5 Styrker og svakheter

De deskriptive observasjonene overnevnt gir indikatorer til styrker og svakheter i datasettet, hvor man finner at hos både kjønn og alder har man god representativitet på spredningen av alder, men også en kjønnsfordeling som er nærmest likestilt. Videre ligger den andre styrken i datasettets utvalg på 2.707.788 (2,7 millioner +). Dette er et viktig aspekt innen statistikk hvor flere observasjoner kan medføre en mer nøyaktig fordeling i forhold til populasjonen man ønsker å generalisere til (Stock & Watson, 2015).

For svakheter er det klar indikasjon på skjevfordeling i kriminalitetsvariabelen, det vises at gjennomsnittet for avgjorte siktelser er veldig lavt, som tilsier at den gjennomsnittlige personen tar ikke på seg en siktelse. Det andre er også øvrekvartilen som indikerer at det er personer observert på ytterkanten av fordelingen og sjeldent oppstår i samfunnet, men som strekker ut gjennomsnittet i datasettet.

Selv om det er visse tegn på svakheter i kriminalitetsvariabelen, anser vi ikke dette som noe som kunne blitt korrigert ved å bruke en annen innsamlingsmetode eller en annen variabeltype. Imidlertid mener vi fortsatt at denne variabelen, til tross for sine svakheter, gir en god representasjon av kriminalitet blant Norges befolkning.

6 Analyse

I denne delen av oppgaven skal vi utforske regresjonsmodellene og sammenhengene disse fanget opp. Det vil gjøres en objektiv tolkning av regresjonsmodellenes resultater, og vil settes i søkelys mot hypotesene vi valgte tidligere. Vi vil først starte med å introdusere regresjonsmodellene vi skal benytte og analysere, for så å svare på antakelsene til regresjon. Deretter vil vi presentere de sentrale funnene til regresjonsmodellene.

6.1 Regresjonsmodellene

For at vi skulle legge til rette for besvarelse av forskningsspørsmålet ble det konstruert fire lineære regresjonsmodeller. Modellene er oppbygd for å undersøke sammenhengen mellom lønn ($\ln_l\text{ønn}$) og de forskjellige uavhengige variablene. Flere av disse uavhengige variablene er kontrollvariabler, også kjent som alder, kjønn og utdanning. Vår primære variabel er kriminalitet og er derfor gjennomgående i alle modellene.

MODELL 1

$$\ln_lønn = \beta_0 + \beta_1 \textit{krim_sum}$$

MODELL 2

$$\ln_lønn = \beta_0 + \beta_1 \textit{krim_sum} + \beta_2 \textit{alder_2021}$$

MODELL 3

$$\ln_lønn = \beta_0 + \beta_1 \textit{krim_sum} + \beta_2 \textit{alder_2021} + \beta_3 \textit{kjønn}$$

MODELL 4

$$\ln_lønn = \beta_0 + \beta_1 \textit{krim_sum} + \beta_2 \textit{alder_2021} + \beta_3 \textit{kjønn} + \beta_4 \textit{gruppert_utdanning}$$

Vår tilnærming til å bygge opp modellene fra 1 til 4 har et formål om at vi skal kunne undersøke sammenhenger og endringer.

I Modell 1 ble det valgt å kun inkludere primærvariabelen kriminalitet. Dette er fordi vi ønsker først og se på hvordan en ekstra siktelse kan påvirke lønnen når vi ikke har kontrollert for, eller fjernet variasjoner som er forklart av andre karakteristikk ved en person som alder, kjønn eller utdanningsbakgrunn.

Fra modell 2 til 4 har vi stegvis inkludert kontrollvariabler, og introdusert en variabel om gangen. På denne måten får vi observert den direkte koeffisienten til kontrollvariabelen, men også hvordan denne påvirker lønn. Vi får også undersøkt endringer i kriminalitets effekt på lønn når en kontrollvariabel om gangen blir inkludert.

6.2 Analyse av regresjonens antakelser

Det første vi skal avklare i analysedelen er de forskjellige antakelsene testet på regresjonsmodellene, hvor vi vil introdusere resultatene fra disse og hva de indikerer.

6.2.1 Linearitet

Den første antakelsen er linearitet. På bakgrunn av microdatas begrensninger mister vi direkte tilgang til individuelle observasjoner som forhindret oss i å utføre relevante tester på linearitet. Blant dette var manglende tilgang på Scatter Plot, hvor den eneste visuelle tilnærmingen vi hadde tilgang på var histogram som ikke var bra nokk til å trekke konklusjoner om linearitet i residualene. Dermed kan vi ikke si noe konkluderende om linearitetsantakelsen.

6.2.2 Heteroskedastisitet

For antakelsen om homoskedastisitet gjennomførtes det en Breusch-pagan test på alle 4 modellene.

TABELL 5 HETEROSKEDASTISITET TESTER PÅ REGRESJONSMODELLERNE

	MODELL 1		MODELL 2		MODELL 3		MODELL 4	
TEST TYPE	chi-verdi	p-verdi	chi-verdi	p-verdi	chi-verdi	p-verdi	chi-verdi	p-verdi
Breusch-pagan	109,9	0	3724,8	0	8696,0	0	748,8	0

Tabell 5 Data uthentet av microdata koden, bestående av resultater funnet ved regresjon. Tabellen er utarbeidet i Excel.

For den første modellen fikk vi en chikvadrat-verdi (chi-verdi) på 109,9 og en tilhørende lav p-verdi. Dette vil si at modellens hypotesetest for eksistensen av homoskedastisitet var negativ, som tilsier at heteroskedastisitet i residualfordelingen er mer sannsynlig. Videre ble det gjennomført tester på modell 2, 3 og 4 som ga tilnærmet like resultater, noe som tilsier at det er signifikante resultater for at heteroskedastisitet eksisterer i residualfordelingene.

6.2.3 Multikollinearitet

Den tredje antakelsen er multikollinearitet hvor vi benyttet VIF Test for å teste om multikollinearitet eksisterer i modellene.

TABELL 6 VIF-TEST PÅ REGRESJONSMODELLER

	MODELL 1	MODELL 2	MODELL 3	MODELL 4
VARIABEL	VIF-verdi	VIF-verdi	VIF-verdi	VIF-verdi
krum_sum	1,00	1,00	1,01	1,02
gruppert_utdanning	X	X	X	1,03
alder_2021	X	1,00	1,00	1,01
kjønn	X	X	1,01	1,03
GJENNOMSNIITT	1,00	1,00	1,01	1,02

Tabell 6 Data uthentet av microdata koden, bestående av resultater funnet ved regresjon. Tabellen er utarbeidet i Excel.

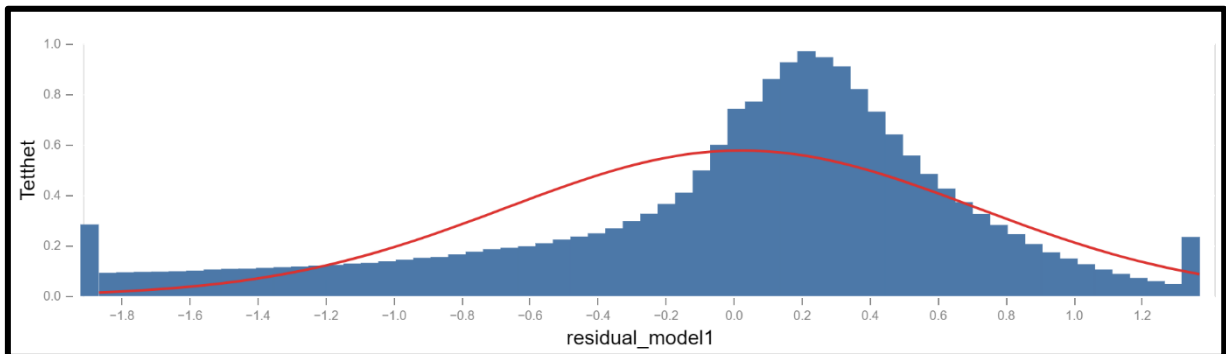
Den første modellen indikerte lav sannsynlighet av multikollinearitet med en gjennomsnittlig VIF- verdi på 1, som tyder på ingen signifikante funn av multikollinearitet. Videre hadde modell 2 like VIF- Verdier, med et gjennomsnitt på 1 som er fortsatt lav. Modellene 3 og 4 hadde også en VIF- verdier som var nær 1, og regnes heller ikke som signifikante.

For å oppsummere viste alle modellene lave VIF- Verdier, og tyder ikke på noe signifikant preg av multikollinearitet. Dette antyder at ingen av verdiene bærer korrelasjoner til hverandre som kan ansees som signifikant.

6.2.4 Normalitet

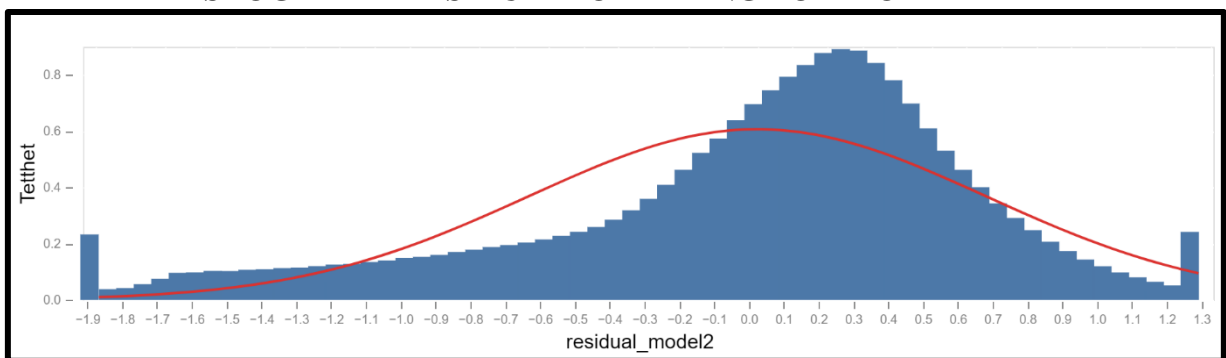
Normalitet er den siste antakelsen som er viktig å bemerke. Det ble utført visuelt histogram med normalfordelingslinje for å sjekke normalitet, samtlig en “jarque-bera” test.

HISTOGRAM 1 RESIDUALFORDELING FOR MODELL 1



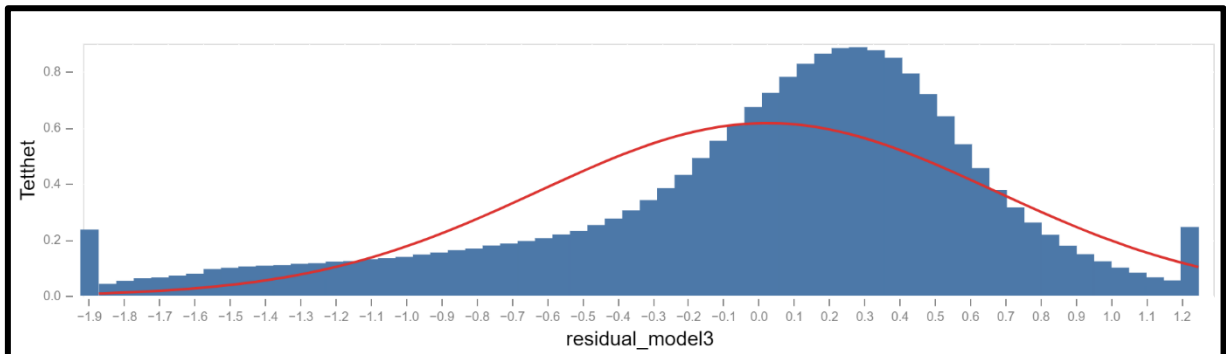
Histogram 1 Oversikt over residualfordeling og normalfordelingslinje utgitt av microdata koden vedlagt. Histogrammet er utarbeidet i Vega Editor.

HISTOGRAM 2 RESIDUALFORDELING FOR MODELL 2



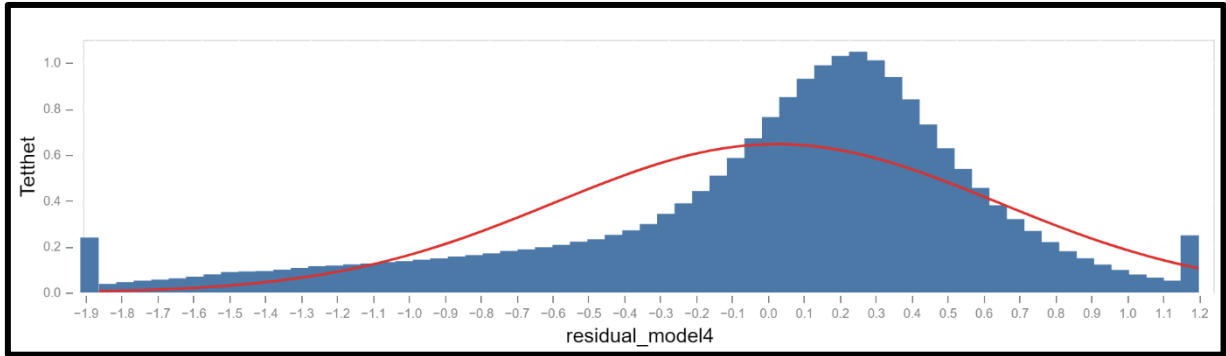
Histogram 2 Oversikt over residualfordeling og normalfordelingslinje utgitt av microdata koden vedlagt. Histogrammet er utarbeidet i Vega Editor.

HISTOGRAM 3 RESIDUALFORDELING FOR MODELL 3



Histogram 3 Oversikt over residualfordeling og normalfordelingslinje utgitt av microdata koden vedlagt. Histogrammet er utarbeidet i Vega Editor.

HISTOGRAM 4 RESIDUALFORDELING FOR MODELL 4



Histogram 4 Oversikt over residualfordeling og normalfordelingslinje utgitt av microdata koden vedlagt. Histogrammet er utarbeidet i Vega Editor.

Fordelingen i histogrammene viser med likhet for alle modellene tilsynelatende en skjevhetfordeling, hvor mye kan virke og være konsentrert rundt gjennomsnittet, men vi ser at de negative verdiene strekker seg lengre enn de positive verdiene for alle modellenes histogrammer. Denne lengre halen av fordelingen kan også bekreftes av en "skjevhet"-test vist i tabellen under. Testen indikerte høyt negative verdier for alle modellene. Dette resultatet tilsier at den venstre siden, altså de negative feilleddene som observeres under 0 strekker seg mye lengre enn feilleddene observert over 0. Hadde denne verdien vært nær 0 ville det ikke vært signifikant forskjell mellom den høyre og venstre siden av fordelingen (Stock, 2020, s. 64).

TABELL 7 NORMALITET TESTER PÅ RESIDUALFORDELINGEN

	MODELL 1		MODELL 2		MODELL 3		MODELL 4	
TEST TYPE	test-verdi	p-verdi	test-verdi	p-verdi	test-verdi	p-verdi	test-verdi	p-verdi
Jarque-Bera	333390,9	0	402443,3	0	493485,8	0	683346,7	0
Skjevhet	-459,3	0	-476,4	0	-510,8	0	-551,4	0

Tabell 7 Oversiktens data kommer fra normaltest på residualfordelingen, gjennomført i microdata koden vedlagt. Tabellen er utarbeidet i Excel.

Histogrammet er derimot bare en visualisering, og "jarque-bera" testen vil kunne gi ytterligere resultater. Denne ble gjennomført på alle modellene, og modell 1 fikk en ekstremt lav p-verdi, som indikerer at testen har avvist hypotesen om en sann normalfordeling, som vil si at residualene ikke følger en normalfordeling. Dette resultatet ble også observert for modell 2, 3 og 4. Dette i kombinasjon med histogrammene og skjevhetstesten tilsier at modellene ikke følger en normalfordeling.

6.2.5 Konsekvenser ved brudd på normalitet & homoskedastisitet

Når residualene er skjevfordelt eller homoskedastiske kan dette påvirke tolkningen av resultatene. Dette er fordi når normalfordeling- og homoskedastisitetsantakelsen ikke er oppfylt kan p-verdiene være mindre nøyaktige, og kan gi indikasjoner til at effekter er mer eller mindre signifikante enn hva dem egentlig er. Denne unøyaktigheten skyldes av den høye variansen til residualene fordi observasjonene på venstresiden til alle modellene ikke er likestilt med den høyre siden av fordelingen. Den samme effekten på p-verdiene oppstår også av de heterogene residualer, ettersom variansen ikke er konstant.

Selv om skjevfordeling og ukonstant varians av residualene kan påvirke p-verdiene, er det viktig å merke seg at dette ikke nødvendigvis påvirker koeffisientene i modellene.

Koeffisientene er fortsatt den samme prediksjonen av sammenhengen mellom variablene, men det er tolkningen av disse koeffisientene som kan være påvirket av feilaktige p-verdier. For å håndtere skjevfordelte eller heterogene residualer, kan det være nyttig å utføre andre analytiske metoder som tar hensyn til dette, men på bakgrunn av nivået for denne oppgaven på bachelornivå har vi valgt å ikke utforske slike justeringer.

6.3 Regresjonsanalyse

Nå som vi har analysert brudd eller avvik på antakelsene ved regresjonsmodellene, går vi videre til å undersøke modellenes direkte resultater.

Ett viktig element som er verdt å påpeke først, er at lønnsvariabelen som skal predikeres kan kun være for året 2021, dette er fordi det kan oppstå endringer i lønnsdannelsen mellom årene som kan endre lønnsobservasjonene i datasettet fra år til år. Det andre er at effekten kriminalitet har på lønn også er en etterkommende effekt på lønn fordi kriminalitet er uthentet fra årsperioden 2002 til 2020, dermed får vi effekten som oppstår etter avgjorte siktelser i perioden.

TABELL 8 REGRESJONSMODELLENES RESULTATER

VARIABEL	LØNN	STANDARDFEI L	P- VERDI	FORKLARINGSGRAD (R ²)	MODEL L
krim_sum	- 0,003743	0,000146	0	0,00024	1
krim_sum	- 0,001638	0,00014	0	0,08271	2
alder_2021	0,01468	0,00003	0		
krim_sum	- 0,005833	0,000138	0	0,10945	3
alder_2021	0,014487	0,000029	0		
kjønn	- 0,230948	0,00081	0		
krim_sum	- 0,001485	0,000132	0	0,18574	4
alder_2021	0,013636	0,000028	0		
kjønn	- 0,281554	0,000781	0		
gruppert_utdannin g	0,177267	0,000352	0		

Tabell 8 Resultatene fra regresjonsmodellene uthentet fra microdata kodens resultat, koden er vedlagt oppgaven. Tabellen er utarbeidet i Excel.

6.3.1 Kriminalitet

Den første modellen, modell 1, fokuserer isolert kun på kriminalitet og viser derfor den direkte korrelasjonen mellom kriminalitet og lønn når vi ikke har kontrollert for andre faktorer. Vi observerer at denne modellen har den laveste forklaringsgraden på 0,024%. Videre fant vi også signifikant sammenheng mellom kriminalitet og lønn, med lave p-verdier. Denne enkle regresjonsmodellen ga resultater som predikerte at med en ekstra avgjort siktelse, eller en enhets økning i kriminalitetsvariabelen, medførtes det en reduksjon i lønn på -0,37%.

6.3.2 Kontrollert for alder

I den andre modellen ble den første kontrollvariabelen, alder, introdusert og alle resultatene fra denne var signifikante.

Denne modellen predikerte at med alderen vil lønnen til en person i gjennomsnitt øke med 1,48%, når kriminalitet holdtes konstant. Vi så også en økning i forklaringsgraden, hvor

denne ble 8,271% når alder ble inkludert. Dette indikerer at alder bidrar i en større grad til å forklare lønnsvariasjoner.

Videre fant vi endring for koeffisienten til kriminalitet fra modell 1 til modell 2, hvor reduksjonen på lønnen nå ble -0,16%, som var mindre negativ i forhold til forrige modell. Dette vil si at deler av forklaringen kriminalitet hadde på lønn var egentlig også forklart av alderen til en person, som vil si at når alder ble inkludert fikk vi et klarere bilde av hvordan kriminalitet faktisk påvirker lønn, uavhengig av alderen deres.

6.3.3 Kontrollert for kjønn

Modell 3 introduserte kjønnsvariabelen, som ikke var inkludert i modell 1 eller 2. Etter denne introduksjonen så vi en endring i kriminalitetskoeffisienten, hvor reduksjonen nå ble noe større på -0,58%. Dette vil si at kjønn er en variabel som har en sammenheng med både kriminalitet og lønn. Vi fant også at kvinner opplevde i gjennomsnitt -20,62% lavere lønn enn menn.

Fra datakapittelet introduserte vi kjønnsforskjellene, hvor 79% av alle avgjorte siktelser i datasettet var menn (Diagram 1). Dette i sammenheng med den lavere lønnen til kvinner utgitt av regresjonsmodellen, kan det indikere at den økte lønnsreduksjonen fra kriminalitetskoeffisienten (-0,58%), stammer fra et utgangspunkt hvor menn opptar større andel av alle avgjorte siktelser, men også tjener mer enn kvinner i gjennomsnitt.

6.3.4 Kontrollert for utdanning

I den siste modellen var alle kontrollvariablene inkludert, men til forskjell fra forrige modell ble nå utdanningsvariabelen introdusert. Denne modellen som alle andre fant signifikante resultater, men var også den modellen med høyest forklaringsgrad på 18,574%.

Når utdanningsnivået ble inkludert observerte vi en mindre negativ kriminalitetskoeffisient. Denne koeffisienten var for modell 4 på -0,15%. Denne effekten er betydelig mindre negativ i forhold til forrige modell (modell 3), som betyr at utdanning var en indirekte effekt vi observerte gjennom kriminalitetskoeffisienten i den forrige modellen. Sett i søkelys av det høye standardavviket vi fant på utdanning i datasettet, kan dette stamme fra personer som har kriminell historikk også har forskjellige utdanningsnivå.

Dermed når utdanning ble inkludert ble denne underliggende effekten kontrollert for, og medførte at kriminalitets effekt på lønn var mer presis. Videre indikerte denne modellen også at med hvert ekstra registrert utdanningsnivå ville det øke lønnen med 19,39% i gjennomsnitt.

6.4 Hypoteseanalyse

Regresjonsanalysen vår fant signifikante resultater i alle modellene. Med p-verdier lavere enn det forhåndsbestemte signifikansnivået på 0,05 fant vi bevis for å avvise nullhypotesene.

HYPOTESE 1

H0: Kriminalitet har ingen signifikant sammenheng med lønn

H1: Kriminalitet har en signifikant sammenheng med lønn

HYPOTESE 2

H0: Kriminalitet har ingen signifikant sammenheng med lønn når kontrollvariablene alder, kjønn og utdanning er inkludert

H1: Kriminalitet har en signifikant sammenheng med lønn når kontrollvariablene alder, kjønn og utdanning er inkludert

I hypotese 1, som tok utgangspunkt i modell 1, fant vi p-verdier lavere enn signifikansnivået og koeffisient som $\neq 0$. Dette tydet det på at vi har tilstrekkelig bevis for å anta at det eksisterer en signifikant sammenheng mellom kriminalitet og lønn, dermed at nullhypotesen (H0) forkastes.

Gjeldende hypotese 2 som tok utgangspunkt i modell 4, tydet analysen på en forkastelse av nullhypotesen (H0) på samme grunnlag som hypotese 1, hvor p-verdiene var lavere enn signifikansnivået. Dette vil si at det er en signifikant sammenheng mellom kriminalitet og lønn, men også kontrollert for kjønn, alder og utdanning.

7 Diskusjon

Fra analysen var modell 4 den som utga de mest presise resultatene siden den inkluderte alle kontrollvariablene, dermed blir denne modellen utgangspunktet for størsteparten av diskusjonsdelen. I dette kapittelet skal vi drøfte resultatene i sammenheng med teoriene fra kapittel 2.

Et sentralt funn var at kriminalitet har en statistisk signifikant betydning på lønn. Denne betydningen var negativ men relativt liten på -0,15%. Dette vil si at for gjennomsnittslønnen i Norge som var på 553.906kr i 2021, tilsvarte det en reduksjon på 831kr. Forklaringer på hvorfor denne endringen er såpass liten og negativ skal vi diskutere i de neste avsnittene.

7.1 Alvorlighetsgrad?

Den lave endringen i lønn knyttet til kriminalitet i 2021 kan først og fremst forklares ved å se på alvorlighetsgraden av kriminelle handlinger. I 2020 utgjorde forenklete forelegg den største delen av straffereaksjonene i Norge, med hele 78% av tilfellene ifølge Statistisk Sentralbyrå (SSB, 2024, Straffereaksjoner).

DIAGRAM 2 OVERSIKT OVER STRAFFEREAKSJONER I 2020

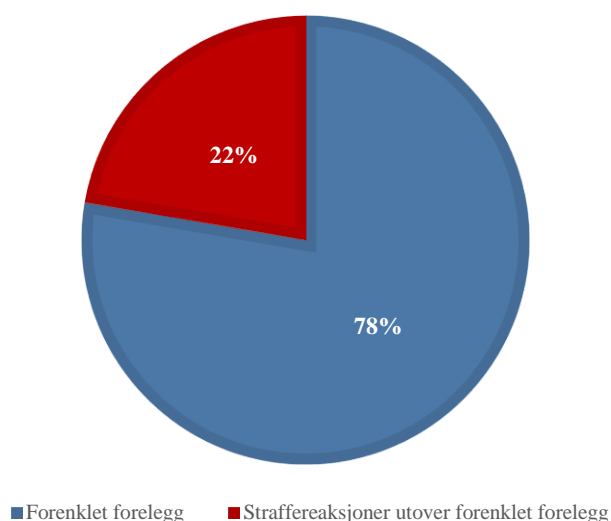


Diagram 2 Data uthentet fra SSB (SSB, 2024, Straffereaksjoner), informasjonen er uthentet fra året 2020 som er gjeldende siste året for variabelen brukt i analysen. Straffereaksjoner utover forenklet forelegg representerer alle andre reaksjoner i statistikken. Diagrammet er utarbeidet i Excel.

Forenklete forelegg er vanligvis ikke assosiert med langsiktige konsekvenser som kan påvirke lønn. Dette skyldes at slike forelegg ikke fører til ytterligere straffereaksjoner utover forelegget som blir gitt på stedet. Dermed blir de heller ikke registrert i bøte- og strafferegisteret (Politiet, 2024a). Denne mangelen på langsiktige konsekvenser kan være en underliggende årsak til den lave effekten som ble observert i analysen. Selv om forenklete forelegg er inkludert som avgjorte siktelsener i datasettet, medfører denne reaksjonen ikke ringvirkninger til begrensninger av politiattest eller en ubetinget dom.

7.1.1 Alvorligere straffereaksjoner

Derimot er det de mer alvorlige straffereaksjonene utover forenklet forelegg som potensielt kan ha en betydelig innvirkning på lønn, og kan forklare hvorfor effekten vi fant var negativ. Ifølge humankapitalteorien kan slike straffereaksjoner medføre tap av humankapital som førerkort, utdanning og yrkesbevilgninger. Dette kan begrense enkeltpersoners jobbmuligheter, spesielt innen yrker som krever spesifikke kvalifikasjoner, godkjenninger eller politiattest. For eksempel kan yrker innen helsevesen, politi og arbeid med barn og unge være sterkt begrenset for personer med en kriminell historikk.

Når det gjelder politiattest kan dette midlertidig eller varig begrense tilgangen til visse utdanninger eller yrker. Men det er viktig å utheve at yrkesfelt som ikke kan stille krav til politiattest, er i flertallet, noe som kan forklare hvorfor den negative lønnsreduksjonen ikke er særlig stor (Politiet, 2024b). Videre er det også begrensninger til hvilke yrkesgrupper som har lovhjemmel for uthenting av denne attesten, samt hvilken type attest de kan be om (Politiet, 2024b). Dette kan bety at arbeidsgiveren ikke alltid har tilgang til den fullstendige straffehistorikken til en jobbsøker, og derfor ikke kan gjøre en fullstendig vurdering av vedkommende. Dermed kan det være at arbeidsgivers begrenset innsyn også kan være en delvis forklaring på den lave lønnsreduksjonen vi fant.

7.1.2 Tapt arbeidserfaring

Videre, gjennom rammeverket for signaleffekten, kan vi forstå ytterligere hvorfor det oppstår en negativ sammenheng mellom kriminalitet og lønn, utover humankapitalteorien. For eksempel kan hull i CV-en som skyldes en betinget dom, altså fengselsstraff, signalisere til arbeidsgiver at jobbsøkeren har en avvikende arbeidshistorikk sammenlignet

med andre i samfunnet. Det er imidlertid viktig å huske at ikke alle med hull i CV-en nødvendigvis har en kriminell historikk.

Samtidig er det også verdt å merke seg at personer med en kriminell historikk kan ha tettet dette hullet mens de sonet sin dom. Dette har litt av sinn argumentasjon innenfor humankapitalteorien. Når en person har blitt utsatt for en av disse strengere straffereaksjonene som medfører ubetinget dom, får de i Norske fengsler like muligheter som resten av samfunnet til videreutdanning. En slik utdanningsmulighet under soning kan bidra til å lukke gapet i CV-en som opprinnelig ville vært der på grunn av en sonet dom. Dermed kan det være at når personer velger å utdanne seg under fengselsopphold, unngår dem de negative konsekvensene av sin kriminelle historikk på CV-en. Dette gjør at arbeidsgivere ikke kan avdekke en tilknyttet kriminalitet, noe som delvis kan forklare hvorfor effekten som ble observert heller ikke var så betydelig stor i Norge.

7.2 Videre forskning

Oppsummerende ser vi at det er flere faktorer i Norges samfunn som kan ha bidratt til de resultatene vi oppnådde. Dette understreker områder som kan være viktige for videre forskning. Ettersom vi erkjenner at vår oppgave bare har skrapet overflaten, anbefaler vi at neste studie fokuserer på lønnsammenhenger mellom ulike straffereaksjoner. Dette kan gi forskeren et klarere bilde av hvordan ulike grupperinger av kriminalitetens alvorlighetsgrad faktisk påvirker lønn i Norge.

Videre oppdaget vi i analysen at utdanning var assosiert med lavere kriminalitetskoeffisient, noe som åpner for en spennende utforskning. Gjennom en kvantitativ interaksjonsanalyse kan man undersøke hvordan lønnen påvirkes av kriminalitet, avhengig av ulike typer utdanningsbakgrunner. Dette er interessant fordi ved å analysere datasettet vårt, har vi fått en oversikt over antallet avgjorte siktelsler fordelt etter utdanningsnivå, som illustrert i grafen nedenfor:

DIAGRAM 3 KRIMINALITET FORDELT PÅ UTDANNINGSNIVÅ

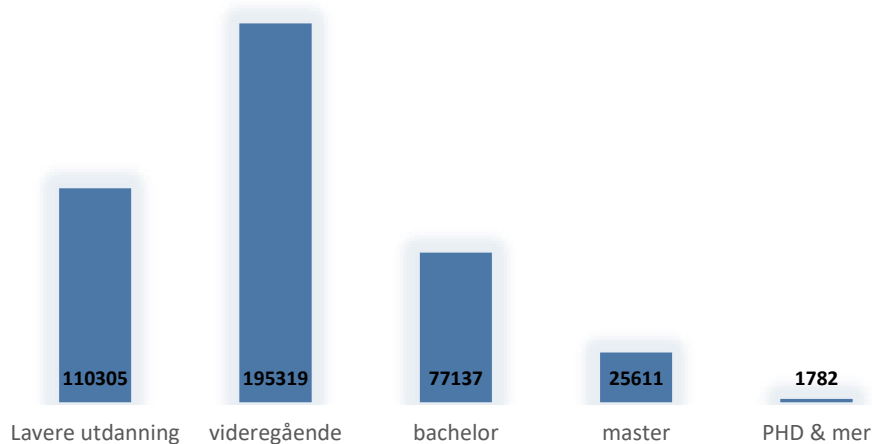


Diagram 3 Viser avgjorte saker fordelt på utdanningsnivå, lavere utdanning er også inkludert i gruppen ingen utdanning. Dataen er uthentet fra koden vedlagt oppgaven og utarbeidet i Excel.

Denne fordelingen understreker at kriminalitet er mest utbredt blant dem med lavere utdanningsnivåer, men også forekommer på høyere nivåer, inkludert ph.d.-nivå.

Resultatene fra en slik studie kan gi en dypere forståelse av hvordan kriminalitet faktisk påvirker lønnen på ulike utdanningsnivåer.

Vi tror at i lys av diskusjonen om politiattester og autorisasjoner, kan høyere utdanning oppleve en betydelig større lønnsreduksjon som følge av kriminalitet enn det vi fant i vår undersøkelse. Dette skyldes at personer med høyere utdanning ofte innehar spesielle autorisasjoner og bevilgninger som personer med lavere utdanningsnivåer ikke har. I tillegg kan høyere utdanning være assosiert med strengere krav til politiattester. Dette kan bety at lønnseffektene fra slike studier vil være forskjellige fra våre funn, da personer med høyere utdanning kan miste mer opparbeidet humankapital enn gjennomsnittspersoner med videregående utdanning i Norge.

8 Konklusjon

Oppgaven har forsøkt å besvare det overordnede forskningsspørsmålet, “Påvirker kriminalitet lønn i Norge?”

Gjennom analysen har vi gjort funn som tyder på en signifikant sammenheng mellom kriminalitet og lønn i Norge for året 2021. Effekten som oppsto var relativt liten, og utgjorde en lav reduksjon i lønn på -0,15% for hver avgjort siktelse. Denne lønnsreduksjonen vi fant var en etterkommende effekt av tidligere kriminalitet fra perioden 2002 til 2020, som dermed kan tilsvare en økonomisk konsekvens fra en tidligere kriminell historikk.

Det var flere faktorer som kan ha medvirket i hvorfor denne reduksjonen ikke var særlig stor. Blant disse faktorene kan bruken av forenklet forelegg som en straffereaksjon være sentral ettersom at det ikke medfører langsiktige konsekvenser. Videre kan effekter av negativ humankapital og signaleffekter i arbeidsmarkedet ha konsekvenser som forklarer hvorfor det oppsto en negativ sammenheng i det hele tatt.

9 Litteraturliste

- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of political economy*, 76(2), 169-217.
- Becker, G. S. (1975). *Human capital: A theoretical and empirical analysis, with special reference to education*. NBER.
- Bhuller, M., & Røgeberg, O. (2022). *Kjennetegn ved barn og unge som begår kriminelle handlinger og virkninger av straff*. Hentet fra: <https://www.regjeringen.no/contentassets/e8f0236bb0784a83911537df3d1fb9f2/rapport-ungdomskriminalitet.pdf>
- Eggen, F. W., Bjørn Gran, M., Norberg-Schulz, & Steen, J. I. (2022). *Samfunnsøkonomiske kostnader av kriminalitet (6-2022)*. Hentet fra: <https://www.regjeringen.no/contentassets/c1e4467f7fea4adf346429cb525b46/r06-2022-samfunnsokonomiske-kostnader-av-kriminalitet.pdf>
- Eirik-Mathias, Rummelhoff, B., & Frøslie, K. F. (2024). Median. Hentet fra: <https://snl.no/median>
- Hartmann, K., Krois, J., & Rudolph, A. (2023). *Statistics and Geodata Analysis using R (SOGA-R)*. Department of Earth Sciences. Hentet fra: <https://www.geo.fu-berlin.de/en/v/soga-r>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.
- Johansen, C. M., & Sundbye, L. M. T. (19.11.2019). *Kvantitative og kvalitative metoder*. Hentet fra: <https://ndla.no/article/20755>
- Kriminalomsorgen. (2024). *Utdanning og opplæring*. Hentet fra: <https://www.kriminalomsorgen.no/utdanning-og-opplaering.527955.no.html>
- Lekve, M., & Arnesen, H. S. (2023). *Hvor likestilt er vi?* Hentet fra: <https://www.ssb.no/befolkning/likestilling/artikler/hvor-likestilt-er-vi>
- Merriam-Webster. (2024). alphanumeric. In *merriam-webster.com dictionary*. Hentet fra: <https://www.merriam-webster.com/dictionary/alphanumeric>
- Microdata. (2024a). *Om microdata.no*. Hentet fra: <https://www.microdata.no/om-microdata-no/>
- Microdata. (2024b). *Variabeloversikt* Hentet fra: <https://microdata.no/discovery/variables?sortBy=NAME&sortType=ASC>
- Politiet. (2024a). *Bøter, forelegg og andre straffereaksjoner*. Lovdata. Hentet fra: <https://www.politiet.no/rad/boter-forelegg-og-andre-straffereaksjoner/#hva-er-et-forelegg-og-forenklet-forelegg-button>
- Politiet. (2024b). *Opplysninger i politiattesten*. Politiregisterloven. Hentet fra: <https://www.politiet.no/tjenester/politiattest/opplysninger-i-politiattesten/#hvor-lenge-er-en-politiattest-gyldig-button>
- Rege, M., Skarðhamar, T., Telle, K., & Votruba, M. (2009). The effect of plant closure on crime. *Statistics Norway, Research Department*(593).
- Sikt, & SSB. (2022). *Brukermanual for microdata.no*. Microdata. Hentet fra: <https://www.microdata.no/wp-content/uploads/2022/04/brukermanual-no-1.pdf>
- Sikt, & SSB. (2024). *Personvernerklæring*. Microdata. Hentet fra: <https://www.microdata.no/personvernerklaering/>
- Skatteetaten. (2024). *Fribeløp for ungdom*. Hentet fra: <https://www.skatteetaten.no/satser/fribelop-for-ungdom/>
- Solberg, R. (2021). *Sammen mot barne-, ungdomsog gjengkriminalitet*. regjeringen.no Hentet fra: <https://www.regjeringen.no/contentassets/d49ad5d168fc46b0bf0f4be0830d2169/no/pdfs/stm202020210034000dddpdfs.pdf>

- Spence, M. (1978). Job market signaling. In *Uncertainty in economics* (pp. 281-306). Elsevier.
- SSB. (2024). *Straffereaksjoner*. Hentet fra: <https://www.ssb.no/statbank/table/13579>
- Stock, J. H., & Watson, M. W. (2015). *Introduction to econometrics* (Updated Third, Global edition. ed.). Pearson.
- Lov om straff (LOV-2005-05-20-28), (2005). Hentet fra: https://lovdata.no/dokument/NL/lov/2005-05-20-28/*#*
- Thadewald, T., & Büning, H. (2007). Jarque–Bera test and its competitors for testing normality—a power comparison. *Journal of applied statistics*, 34(1), 87-105.
- Ubøe, J. (2017). *Introductory Statistics for Business and Economics: Theory, Exercises and Solutions* (1st 2017. ed.). Cham: Springer Nature. Hentet fra: <https://doi.org/10.1007/978-3-319-70936-9>
- Wikipedia. (2024). *Kriminalitet*. Hentet fra: <https://no.wikipedia.org/wiki/Kriminalitet>
- Wooldridge, J. M. (2020). *Introductory econometrics : a modern approach* (Seventh edition. ed.). Cengage Learning.

10 Microdata kode (Vedlegg)

Undernevnt ligger koden for microdata, hvor dette er den endelige brukt i oppgaven. Alle tabeller og visuelle grafer gjennomgått i oppgaven uten kilder er basert på denne, og resultatene fra den.

START KODE

```
//Koble til databank
require no.ssb.fdb:26 as db

//Lage dataset
create-dataset lønn_krim

//Import
import db/INNTEKT_WLONN 2021-12-31 as inntekt_2021
import db/NUDB_BU 2021-12-31 as utd
import db/BEFOLKNING_FOEDSELS_AAR_MND as fødmndår
import db/BEFOLKNING_KJOENN as kjønn

//Gruppere kriminalitet i en summert variabel og import
import db/KRIMINALITET_SIKTEDE_LOVANT 2020-12-31 as kriminalitet_avgjort_2020
import db/KRIMINALITET_SIKTEDE_LOVANT 2019-12-31 as kriminalitet_avgjort_2019
import db/KRIMINALITET_SIKTEDE_LOVANT 2018-12-31 as kriminalitet_avgjort_2018
import db/KRIMINALITET_SIKTEDE_LOVANT 2017-12-31 as kriminalitet_avgjort_2017
import db/KRIMINALITET_SIKTEDE_LOVANT 2016-12-31 as kriminalitet_avgjort_2016
import db/KRIMINALITET_SIKTEDE_LOVANT 2015-12-31 as kriminalitet_avgjort_2015
import db/KRIMINALITET_SIKTEDE_LOVANT 2014-12-31 as kriminalitet_avgjort_2014
import db/KRIMINALITET_SIKTEDE_LOVANT 2013-12-31 as kriminalitet_avgjort_2013
import db/KRIMINALITET_SIKTEDE_LOVANT 2012-12-31 as kriminalitet_avgjort_2012
import db/KRIMINALITET_SIKTEDE_LOVANT 2011-12-31 as kriminalitet_avgjort_2011
import db/KRIMINALITET_SIKTEDE_LOVANT 2010-12-31 as kriminalitet_avgjort_2010
import db/KRIMINALITET_SIKTEDE_LOVANT 2009-12-31 as kriminalitet_avgjort_2009
import db/KRIMINALITET_SIKTEDE_LOVANT 2008-12-31 as kriminalitet_avgjort_2008
import db/KRIMINALITET_SIKTEDE_LOVANT 2007-12-31 as kriminalitet_avgjort_2007
import db/KRIMINALITET_SIKTEDE_LOVANT 2006-12-31 as kriminalitet_avgjort_2006
import db/KRIMINALITET_SIKTEDE_LOVANT 2005-12-31 as kriminalitet_avgjort_2005
```

```

import db/KRIMINALITET_SIKTEDE_LOVANT 2004-12-31 as kriminalitet_avgjort_2004
import db/KRIMINALITET_SIKTEDE_LOVANT 2003-12-31 as kriminalitet_avgjort_2003
import db/KRIMINALITET_SIKTEDE_LOVANT 2002-12-31 as kriminalitet_avgjort_2002

recode kriminalitet_avgjort_2020 kriminalitet_avgjort_2019 kriminalitet_avgjort_2018
kriminalitet_avgjort_2017 kriminalitet_avgjort_2016 kriminalitet_avgjort_2015 kriminalitet_avgjort_2014
kriminalitet_avgjort_2013 kriminalitet_avgjort_2012 kriminalitet_avgjort_2011 kriminalitet_avgjort_2010
kriminalitet_avgjort_2009 kriminalitet_avgjort_2008 kriminalitet_avgjort_2007 kriminalitet_avgjort_2006
kriminalitet_avgjort_2005 kriminalitet_avgjort_2004 kriminalitet_avgjort_2003 kriminalitet_avgjort_2002
(missing = 0)

generate krim_sum = kriminalitet_avgjort_2020 + kriminalitet_avgjort_2019 + kriminalitet_avgjort_2018 +
kriminalitet_avgjort_2017 + kriminalitet_avgjort_2016 + kriminalitet_avgjort_2015 +
kriminalitet_avgjort_2014 + kriminalitet_avgjort_2013 + kriminalitet_avgjort_2012 +
kriminalitet_avgjort_2011 + kriminalitet_avgjort_2010 + kriminalitet_avgjort_2009 +
kriminalitet_avgjort_2008 + kriminalitet_avgjort_2007 + kriminalitet_avgjort_2006 +
kriminalitet_avgjort_2005 + kriminalitet_avgjort_2004 + kriminalitet_avgjort_2003 +
kriminalitet_avgjort_2002

//alder variabel generering
generate alder_2021 = 2021 - int(fødmndår/100)

//Konvertere kjønn fra alfanumerisk til numerisk
destring kjønn

//Lage lønnsvariabel
generate lønn = inntekt_2021
keep if inntekt_2021 >= 60000 & alder_2021 >= 15
generate ln_lønn = ln(lønn)

//Begrense utdanning til nivåer og grupperinger og legger det inn i datasettet
generate gruppert_utdanning = 0
replace gruppert_utdanning = 1 if utd == '099901'
replace gruppert_utdanning = 2 if inrange(utd,'099902','301110')
replace gruppert_utdanning = 3 if inrange(utd,'301111','611115')
replace gruppert_utdanning = 4 if inrange(utd,'611116','701999')
replace gruppert_utdanning = 5 if inrange(utd,'711101','801999')
replace gruppert_utdanning = 6 if inrange(utd,'811101','999999')

```

```

//Angir navn til forskjellige grupperinger utdanning
define-labels utdanning_label 0 'Ukjent avvik' 1 'Ingen' 2 'forskole' 3 'videregående' 4 'bachelor' 5 'master' 6
'PHD & mer'
assign-labels gruppert_utdanning utdanning_label

textblock
Normaltest test alle variabler
-----

endblock

correlate krim_sum alder_2021 gruppert_utdanning inntekt_2021 kjønn, obs sig
normaltest krim_sum
normaltest alder_2021
normaltest gruppert_utdanning
normaltest ln_lønn

textblock
Histogrammer
-----

endblock

histogram krim_sum
histogram inntekt_2021
histogram lønn
histogram ln_lønn
histogram ln_lønn, normal
histogram alder_2021
histogram kjønn

textblock
Summarize alle samlet
-----

endblock

summarize alder_2021 krim_sum gruppert_utdanning ln_lønn kjønn kriminalitet_avgjort_2020
kriminalitet_avgjort_2019 kriminalitet_avgjort_2018 kriminalitet_avgjort_2017 kriminalitet_avgjort_2016
kriminalitet_avgjort_2015 kriminalitet_avgjort_2014 kriminalitet_avgjort_2013 kriminalitet_avgjort_2012
kriminalitet_avgjort_2011 kriminalitet_avgjort_2010 kriminalitet_avgjort_2009 kriminalitet_avgjort_2008

```

kriminalitet_avgjort_2007 kriminalitet_avgjort_2006 kriminalitet_avgjort_2005 kriminalitet_avgjort_2004
kriminalitet_avgjort_2003 kriminalitet_avgjort_2002

textblock

Tabeller

endblock

//tabulate krim_gruppert

tabulate gruppert_utdanning

tabulate kjønn

//Regresjon

textblock

Modell 1

endblock

regress ln_lønn krim_sum, ov vif het_bp het_iid het_fstat

textblock

Modell 2

endblock

regress ln_lønn krim_sum alder_2021, ov vif het_bp het_iid het_fstat

textblock

Modell 3

endblock

regress ln_lønn krim_sum alder_2021 kjønn, ov vif het_bp het_iid het_fstat

textblock

Modell 4

endblock

regress ln_lønn krim_sum alder_2021 kjønn gruppert_utdanning, ov vif het_bp het_iid het_fstat

textblock

Residualvariabel, viser residualfordelingen

endblock

ivregress-predict ln_lønn krim_sum, residuals(residual_model1)

ivregress-predict ln_lønn krim_sum alder_2021, residuals(residual_model2)

ivregress-predict ln_lønn krim_sum alder_2021 kjønn, residuals(residual_model3)

ivregress-predict ln_lønn krim_sum alder_2021 kjønn gruppert_utdanning, residuals(residual_model4)

textblock

Residual histogram & normaltest modell 1

endblock

histogram residual_model1

histogram residual_model1, normal

normaltest residual_model1

textblock

Residual histogram & normaltest modell 2

endblock

histogram residual_model2

histogram residual_model2, normal

normaltest residual_model2

textblock

Residual histogram & normaltest modell 3

endblock

histogram residual_model3

histogram residual_model3, normal

normaltest residual_model3

textblock

Residual histogram & normaltest modell 4

endblock

histogram residual_model4

histogram residual_model4, normal

normaltest residual_model4

textblock

Andel kjønn kriminelle

endblock

generate kjønn_kriminalitet = kjønn if krim_sum > 0

histogram kjønn_kriminalitet

tabulate kjønn_kriminalitet

//Krim fordelt på utdanning

generate utd_krim = 0 if gruppert_uttanning == 0 & krim_sum > 0

replace utd_krim = 1 if gruppert_uttanning == 1 & krim_sum > 0

replace utd_krim = 2 if gruppert_uttanning == 2 & krim_sum > 0

replace utd_krim = 3 if gruppert_uttanning == 3 & krim_sum > 0

replace utd_krim = 4 if gruppert_uttanning == 4 & krim_sum > 0

replace utd_krim = 5 if gruppert_uttanning == 5 & krim_sum > 0

replace utd_krim = 6 if gruppert_uttanning == 6 & krim_sum > 0

textblock

Utdanning med avgjort siktelse større en 0

endblock

define-labels utd_krim_label 0 'Ukjent avvik & krim' 1 'Ingen & krim' 2 'forskole & krim' 3 'videregående & krim' 4 'bachelor & krim' 5 'master & krim' 6 'PHD & mer & krim'

assign-labels utd_krim utd_krim_label

tabulate utd_krim

tabulate gruppert_uttanning

summarize gruppert_uttanning utd_krim

SLUTT KODE