Metodevedlegg for rapporten: Covid-19 blant personer født utenfor Norge, justert for yrke, trangboddhet, medisinsk risikogruppe, utdanning og inntekt

FHI har mottatt enkelte oppfølgingsspørsmål av metodisk art til vår rapport (Covid-19 blant personer født utenfor Norge, justert for yrke, trangboddhet, medisinsk risikogruppe, utdanning og inntekt). Vi har derfor valgt å publisere en noe mer inngående metoderedegjørelse der vi også gjør en analyse av betydningen av husholdningsstørrelse og en analyse med alternativ modell (multiplikativ i tillegg til additiv modell). Hovedkonklusjonen er at betydningen av kjønn, alder og bosted blir noe større ved analyser med multiplikativ modell, mens betydningen av trangboddhet forblir begrenset uavhengig av modell. Betydningen av husholdningsstørrelse som tillegg til trangboddhet synes å være beskjeden – men en viss betydning finner vi. Hovedbudskapene i rapporten endres ikke av disse variasjonene i modellspesifikasjoner.

1. Modell

Som redegjort for i rapporten estimerer vi følgende lineære sannsynlighetsmodell (med minste kvadraters metode):

$$y_i = country_i \beta_k^{country} + controls_i^k \beta_k^{controls} + \varepsilon_i \quad (1)$$

der y_i er en indikatorvariabel lik 1 hvis individ i har testet positivt for covid-19/ blitt innlagt med covid-19, og 0 ellers. $country_i$ er en vektor med indikatorvariabler for fødeland (referansekategori Norge), og $controls_i$ er en vektor med justeringsvariablene.

Tabell A: kovariater

k	**Modell**	**Kontrollvariabler**
1	Ujustert	Ingen
2	Demografi	Kjønn og alder (dummy-kodet i 5-års kategorier)
3	Bosted	Kjønn, alder og bostedskommune
4	Yrke	Kjønn, alder, bostedskommune, yrke (2-sifret), næring (2-sifret)
5	Trangboddhet	Kjønn, alder, bostedskommune, trangboddhet
6	Medisinske risikogrupper	Kjønn, alder, bostedskommune, medisinske risikogrupper (14 diagnosegrupper)
7	Utdanning	Kjønn, alder, bostedskommune, høyeste fullførte utdanning (grunnskole, videregående- og fagskole, kort og lang høyskole utdanning, ingen eller ukjent utdanning, personer under 25 år)
8	Inntekt	Kjønn, alder, bostedskommune, husholdningsinntekt (desiler av inntektsfordelingen, husholdningsinntekt per forbruksenhet, justert etter EU-skalaen)
9	Samlet	Kjønn, alder (5-års kategorier), bostedskommune, yrke, trangboddhet, medisinske risikogrupper, utdanning, inntekt

For hver av disse spesifikasjonene estimerer vi en vektor med koeffisienter for (grupper av) fødeland.

Disse har tolkning som forventet forskjell i covid-19-utfall mellom de ulike fødelandene og personer født i Norge, betinget på eventuelle kovariater i modellen. For fødeland $countryC$:

$$\beta_k^{countryC} = E\left[y\middle|country = countryC, \overline{controls_k}\right] - E[y|country = Norway, \overline{controls}_k]$$

Estimatene angir dermed avviket i prosentpoeng i forhold til snittet i referansegruppen (typisk dem født i Norge), og i figurene har vi angitt disse avvikene rundt gjennomsnittet for utfallsvariabelen blant dem født i Norge. For å se på relative endringer, ikke bare absolutte, dividerer vi flere steder i teksten hvert av disse estimatene med gjennomsnittlig forekomst av covid-19 for norskfødte personer, $\bar{y}_{Norway}$.

$$\tilde{\beta}_k^{\ country} = \frac{\beta_k^{\ country}}{\bar{y}_{Norway}}$$

$\tilde{\beta}_k^{\ country}$ har dermed tolkningen som prosentvis forskjell i forventet smitte mellom fødeland (justert for de angitte observerte forskjellene).

2. Lineære og ikke-lineære modeller

I våre analyser har vi primært brukt de estimerte parameterne fra ligning (1) til å studere de marginale effektene av fødeland ($\beta_k^{countryC}$). Vi antar at $E[\varepsilon_i|country_i, controls_i] = 0$, dermed kan vi uttrykke forventede utfall som en funksjon av fødeland og kovariater:

$$E[y|country, controls] = country\beta_k^{\ country} + controls^k\beta_k^{controls}$$

Siden y_i her er en binær variabel, som tar verdien 1 hvis hendelsen (smitte, innleggelse) skjer, og 0 ellers, er sannsynligheten for at hendelsen skjer det samme som forventningsverdien til y. Dermed har vi i vår foretrukne modell følgende uttrykk for sannsynligheten for utfallene som funksjon av fødeland og kovariater:

$$Pr[y = 1|country, controls] = country\beta_k^{\ country} + controls^k\beta_k^{controls}$$

Modellen antar med andre ord en lineær, additiv sammenheng mellom kovariatene og sannsynligheten for smitte og innleggelse. Hvis den underliggende sammenhengen mellom observerbare kjennetegn og forventede utfall er (tilnærmet) lineær, vil en lineær regresjon gi (tilnærmet) riktige marginale effekter. Dersom den underliggende modellen ikke er lineær, vil vår modell være feilspesifisert. Motsatt vil en modell som antar en proporsjonal sammenheng mellom kovariater og forventede utfall, være feilspesifisert dersom den sanne modellen er additiv.

Innenfor anvendt forskning har det gjennom flere tiår pågått en debatt om de ulike fordelene og ulempene ved lineære sannsynlighetsmodeller (minste kvadraters metode) vs ikke-lineære modeller (for eksempel generaliserte lineære modeller med for eksempel logit link funksjon) når den sanne modellen ikke er kjent. Angrist og Pischke (2009) trekkes ofte fram som forsvarere av minste kvadraters metode framfor ikke-lineære modeller. Det er mange argumenter i debatten, men Angrist og Pischkes overordnede poeng synes å være at det eneste vi vet sikkert i anvendte analyser, er at modellen alltid er en forenkling og dermed feilspesifisert. Å ikke ta hensyn til at utfallsvariabelen er binær er ett argument mot minste kvadraters metode, men fordi vi vet at alle forutsetningene bak en logistisk modell heller ikke holder, kan det godt hende at vi lærer mer om underliggende empiriske

mønstre av å bruke minste kvadraters metode framfor en logistisk modell (eller andre ikke-lineære spesifikasjoner). Til syvende og siste er det mange argumenter for og mot begge modelltyper, og valget av modell handler ofte om tradisjoner i faget og hvordan man mener analysenes styrker og svakheter best kan formidles på transparent måte til målgruppa. Heldigvis er det også velkjent at i anvendte analyser endres hovedresultatene sjelden mye av om man velger den ene eller andre modellvarianten, andre typer modelleringsvalg er ofte langt viktigere (se også Hellevik 2009).

I vår rapport har vi ikke klare teoretiske holdepunkter for å anta en spesifikk funksjonsform mellom kovariater og forventede utfall. Som en robusthetstest har vi derfor også estimert en logistisk regresjonsmodell. Denne modellen hører inn under klassen av generaliserte lineære modeller

$$Pr[y = 1|country, controls] = G(country\beta_k{}^{country} + controls^k\beta_k^{controls})$$

I en generalisert lineær modell vil G(z) typisk være en ikke-lineær funksjon som sikrer at de predikerte sannsynlighetene er mellom 0 og 1. I logistisk regresjon antas en logit link, det vil si at G(z) er den logistiske funksjonen

$$G(z) = \frac{exp(z)}{1 - exp(z)}$$

Denne modellen forutsetter at kovariatene medfører proporsjonale skift i odds (p/(1-p)).

2. Resultatenes sensitivitet for modellvalg og husholdningsstørrelse

Målet med modelleringen er å estimere andel smittede etter fødeland, kontrollert for mulige forskjeller i sammensetningen mellom gruppene. Vi finner at de to modellene gir substansielt samme resultater. Data viser at norskfødte har en smitteandel på 1176 per 100 000, og pakistanskfødte en smitteandel på 9174 per 100 000. Dette tilsvarer en odds ratio på (0,09174/(1-0,09174))/(00,01176/(1-0,01176)) 8,5 og en relativ risiko på (0,09174/0,01176) 7,8. I en modell hvor vi justerer for kontrollvariabler må vi fiksere kontrollvariablene på spesifikke verdier for å få andel smittede, f.eks. i en lineær modell justert for alder, kjønn, fylke, trangboddhet og husholdningsstørrelse, fiksert på alder 50-54 år, mann, Oslo, ikke-trangbodd, husholdningsstørrelse 2, er smitteandelen 2701 for norskfødte og 9197 for pakistanskfødte. Dette gir odds ratio 3,6 og relativ risiko 3,4. Tilsvarende justert logistisk modell gir odds ratio 3,8, og predikerte andeler på 8156 for pakistanskfødte og 2284 for norskfødte, fiksert på de samme verdiene som i den lineære modellen. Relativ andel kan vi regne ut til (8156/2284) 3,6.

Det er altså ikke helt trivielt å sammenlikne funnene mellom de to modelltypene, fordi det må gjøres ulike forutsetninger for å komme fra marginaleffekter i lineære modeller til oddsratioer eller relative andeler, og tilsvarende i ikke-lineære modeller for å komme fra oddsratioer til marginaleffekter eller relative andeler. I figur V1 – som er analog til figur 20a i rapporten - har vi valgt å ikke pålegge slike forutsetninger, og dermed bare presentere marginaleffekter for den lineære modellen (minste kvadraters metode) og oddsratioer for den ikke lineære modellen (logistisk modell). For at estimeringen av den ikke-lineære modellen skal gå fortere og for å unngå tomme celler, har vi her – i motsetning til i rapporten – justert for bostedsfylker i stedet for bostedskommuner. Vi ser fra figur V1 at inklusjon av kontrollvariablene reduserer forskjellene mellom landgruppene. Særlig gjelder det når vi justerer for alder, kjønn og bostedsfylke i den ikke-lineære modellen, men i mer begrenset grad når vi justerer for de andre variablene. Store fødelandsforskjeller i smitte gjenstår i begge

modellene etter at vi har justert for alle de observerte variablene samlet (inklusive husholdningsstørrelse).

I rapportens diskusjonsavsnitt drøfter vi hvordan trangboddhet nok ikke fanger de viktigste kjennetegnene ved husholdningene som påvirker smitte, og husholdningsstørrelse er blitt trukket fram som en viktigere variabel. I figur V1 har vi derfor også lagt til husholdningsstørrelse som en variabel i modellene. Vi har kontrollert for antall i husholdningen ved en indikatorvariabel med verdiene 1, 2, 3, 4, 5, 6+. Vi ser fra figuren at dette i liten grad påvirker resultatene.

Referanser

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Hellevik, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity* 43.1 59-74.

Figur V1. Forskjeller i andel covid-19 smitte etter fødeland med norskfødte som referanse. Modellert med lineær regresjon (venstre panel) og med logistisk regresjon (høyre panel). Ujustert modell, samt justerte modeller. Snitt for smittede blant norskfødte er 1176.