Tidsskriftet
DEN NORSKE LEGEFORENING

Små data – (for) store konklusjoner?

KOMMENTAR

CHRISTER THRANE
E-post: christer.thrane@inn.no
Christer Thrane er sosiolog og professor ved Høgskolen i Innlandet.
Forfatteren har ikke oppgitt noen interessekonflikter.

Det er viktig å kartlegge hva som kan forklare sosiale helseforskjeller. Derfor er det fortjenestefullt at Søegaard og Kan bringer denne tematikken over på utbredelsen av covid-19-smitte i Oslo (1). I en analyse basert på aggregerte tall for Oslos 15 bydeler, finner de sterke korrelasjoner mellom tre uavhengige variabler og smittegrad: (A) Bydeler med en høyere andel av innvandrere har *mer* smitte enn bydeler med en lavere andel innvandrere. (B) Bydeler med en høyere andel av folk med høy sosioøkonomisk status har *mindre* smitte enn bydeler med en lavere andel av folk med høy sosioøkonomisk status. (C) Bydeler med en høyere andel av folk som bor trangbodd, har *mer* smitte enn bydeler med en lavere andel av folk som bor trangbodd.

Så gjør forfatterne en multippel regresjonsanalyse, uten at motivet er eksplisitt. Jeg antar de ønsket å si noe om de relative effektene av de nevnte uavhengige variablene, siden senere kommentarer og konklusjon er at kun innvandrerandel har en signifikant effekt i den multiple analysen. Denne multiple regresjonen burde imidlertid ikke vært rapportert, siden konklusjonene fra denne i beste fall er tvilsomme og i verste fall er direkte gale.

Grunnen er at forfatterne ikke har data som kan belegge en slik konklusjon. Her er det flere momenter: (A) Analysen er basert på aggregerte tall for 15 bydeler. Det finnes ingen enighet i litteraturen om hvor mange observasjoner (her: bydeler) som trengs for å gjøre en multippel regresjon med tre uavhengige variabler. Men flere anbefalinger peker mot minst 10–20 observasjoner per uavhengige variabel, noe som i dette tilfellet minimum tilsier 30 observasjoner (2). En annen anbefaling er 50 + 8 observasjoner per uavhengige variabel, dvs. 74 som et minimum (3). (B) Vel så viktig er at de uavhengige variablene ikke bør være sterkt korrelerte, altså at det foreligger multikollinaritet. Søegaard og Kan oppgir korrelasjonene (innvandrerandel vs. sosioøkonomisk status = –0,916; innvandrerandel vs. husstandstetthet = 0,948; sosioøkonomisk status vs. husstandstetthet = –0,883), men uten å trekke konsekvensen av dette, dvs. å legge bort tanken om å rapportere en multippel regresjon. (C) Få observasjoner forsterker problemet med multikollinaritet (4). I sum gir A–C at koeffisientene til forfatternes multiple regresjon ikke er til å stole på og at man heller ikke kan vektlegge dens *p*-verdier.

Søegaard og Kans bruk av statistisk signifikans kan også bemerkes. Bydelsdataene er en populasjon heller enn et tilfeldig utvalg. Signifikansvurderinger kan da forsvares ved å se på dataene som et tilfeldig utvalg fra en *tenkt* superpopulasjon. Men hva denne eventuelt skal være, er uklart. Alternativt kan signifikansvurderinger søkes i modellbasert statistisk teori

(5), men det er ingen spor etter dette. Dermed får den statistiske analysen et skjær av øvelsen «å late som man har et utvalg fordi man ønsker å benytte signifikanstester».

Jeg har delvis reanalysert dataene fra Søegaard og Kan (appendiks 1). To av variablene forfatterne benytter, finnes som rådata i deres tabell 1, sammen med variablene gjennomsnittlig inntekt og andel trangbodde i bydelene. Jeg bruker de sistnevnte som proksier for henholdsvis sosioøkonomisk status og husstandstetthet (forfatterne bruker her to indekser som ikke finnes i deres tabell 1). Denne forskjellen spiller liten rolle, siden det er en høy korrelasjon mellom indeksene og mine proksier (appendiks 1). Tabell 1 i appendikset viser tre bivariate regresjoner mellom smittegrad og de tre uavhengige variablene samt en multippel regresjonsanalyse.

Resultatene for panel A i tabell 1 forteller at en høyere innvandrerandel i bydel samvarierer sterkt positivt med smittegrad for bydel, som vist i figur 1 (appendiks 1). Panel B i tabell 1 viser en negativ samvariasjon mellom smittegrad og gjennomsnittsinntekt (b = –3,36), mens panel C viser en positiv samvariasjon (b = 144,50) mellom smittegrad og andel trangbodde.

Panel D er den multiple regresjonen. Jeg finner omtrent det samme som Søegaard og Kan, men det er flere symptomer på multikollinaritet: (A) Korrelasjonene blant mine uavhengige variabler ligger i intervallet –0,81 til 0,92 (ikke vist), dvs. de er lavere enn hos Søegaard og Kan. Multikollinaritetsproblemet er derfor større hos dem enn hos meg. (B) Standardfeilene er inflaterte i den multiple regresjonen, for innvandrerkoeffisienten øker den fra 2,77 til 7,04, mens den tilsvarende økningen er fra 19,76 til 40,62 for trangboddkoeffisienten. (C) Koeffisienten for trangbodd går fra å være sterkt positiv (144,5) til å bli markant negativ (–36,2), noe som er urimelig. (D) VIF-verdiene, målet på grad av multikollinaritet, indikerer nettopp dette med verdier over 2,5, 5 og 10. Ingen er enige om hva som sikkert definerer multikollinaritet, og alle nevnte terskelverdier benyttes i litteraturen. De fleste vil imidlertid si at terskelen flyttes nedover ved få observasjoner. I sum peker A–D mot multikollinaritet, noe som sammen med de for få observasjonene gjør at vi ikke kan stole på at koeffisientene er korrekte uttrykk for de uavhengige variablenes relative effekter, alt annet likt. Dette rammer i større grad Søegaard og Kans analyse enn min, siden korrelasjonene blant deres uavhengige variabler er større enn blant mine.

Vi bør ikke være redde for å belyse ubehagelige sannheter på sykdoms- og helsefeltet. Søegaard og Kans analyser, og min delvise replikasjon av disse, tilsier at smittegradsvariasjonen for covid-19 blant Oslos bydeler samvarierer med andel innvandrere, sosioøkonomisk status og husstandstetthet. Men ikke mer. Å forsøke å rangere deres relative forklaringskraft er dømt til å mislykkes i en analyse med 15 observasjoner. Innvandrerandel kan være viktigere enn sosioøkonomisk status og husstandstetthet for å forklare forskjellene i smitteandel mellom bydelene i Oslo (6), slik Søegaard og Kans konklusjon mer enn antyder. Men dette trenger vi altså større og bedre data for å kunne belegge enn det forfatterne har.

LITTERATUR:

1. Søegaard EGI, Kan Z et al. Koronasmitte i Oslos bydeler. Tidsskr Nor Legeforen 2021; 141. doi: 10.4045/tidsskr.20.1022. [PubMed][CrossRef]

2. Harrell FE Jr. Regression Modeling Strategies. New York, NY: Springer Forlag, 2001.

3. Tabachnick BG, Fidell LS. Using Multivariate Statistics. 5. utg. Boston, MA: Pearson Education Inc, 2007.

4. Allison PD. Multiple Regression. A Primer. Thousand Oaks, CA: Pine Forge Press, 1999.

5. Aaberge R, Laake P. Om statistiske teoriar for tolking av data. Tidsskr Samfunnsforsk 1984; 25: 156–86.

6. Kjøllesdal M, Indset T, Arnesen T. Covid-19 og innvandrere: Hva sier tallene? Forskersonen 28.3.2021. https://forskersonen.no/covid19-innvandring-kronikk/covid-19-og-innvandrere-hva-sier-tallene/183525

Publisert: 25. mai 2021. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.21.0342