Tidsskriftet
DEN NORSKE LEGEFORENING

Multippel imputering av manglende data

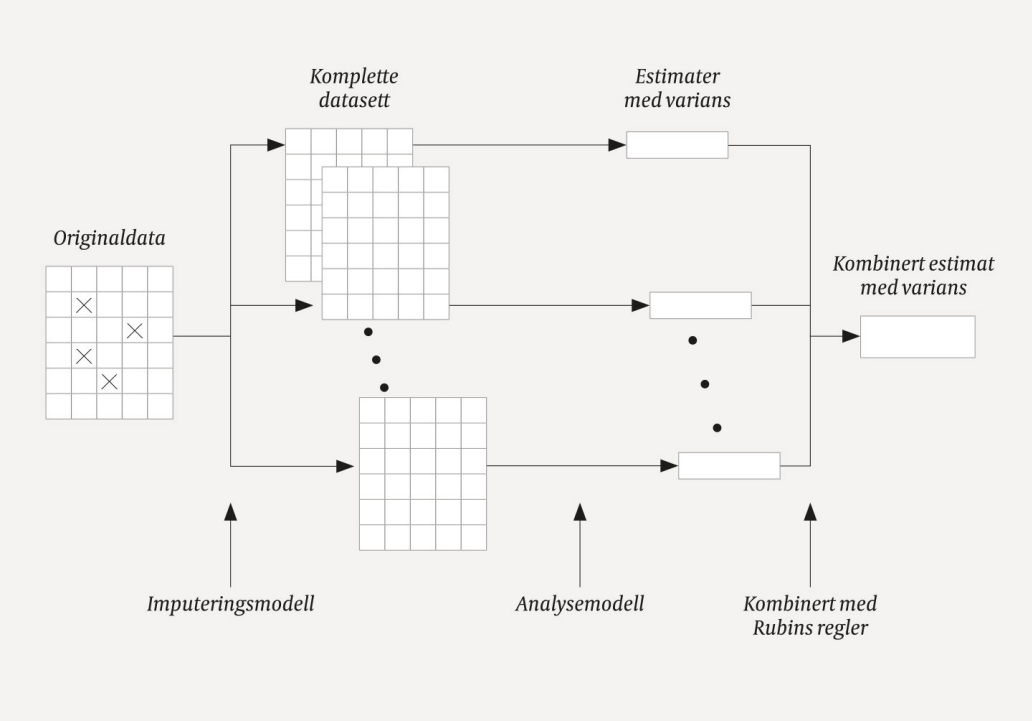
MEDISIN OG TALL

STIAN LYDERSEN

stian.lydersen@ntnu.no
Stian Lydersen er dr.ing. og professor i medisinsk statistikk ved Regionalt kunnskapssenter for barn og unge – psykisk helse og barnevern (RKBU Midt-Norge) ved Institutt for psykisk helse, NTNU.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

De fleste statistiske analysemetoder krever komplette datasett, men i nesten alle studier mangler det enkelte verdier. Multippel imputering er en metode som kan brukes til å håndtere dette.

La oss ta utgangspunkt i et datasett der hver rad i regnearket inneholder data fra en deltaker i en studie. En vanlig situasjon er at det mangler enkelte verdier i datasettet, illustrert med kryss i figur 1. For eksempel kan noen deltakere mangle data om vekt, andre om fysisk aktivitet og andre om samboerstatus. Alle disse inngår som uavhengige variabler i *analysemodellen*, som i dette tilfellet er en regresjonsmodell med blodtrykk som avhengig variabel. Hvis vi analyserer datasettet uten å ta hensyn til manglende data, vil de fleste analysemetoder bare inkludere deltakerne med komplette data, dvs. vi gjør det som på engelsk kalles *complete case analysis*. Dette er sjelden et problem hvis få av deltakerne mangler data, f.eks. under 5 % eller 10 %, men ellers vil den statistiske styrken bli vesentlig redusert på grunn av redusert utvalgsstørrelse. Et mer alvorlig problem er at en slik analyse blir forventningsrett (*unbiased*) bare hvis data mangler helt tilfeldig (*missing completely at random*, MCAR). Analyse basert på multippel imputering, derimot, vil være forventningsrett også hvis data mangler betinget tilfeldig (*missing at random*, MAR) (1).



Figur 1 Fremgangsmåte ved multippel imputering.

Imputeringsmodellen

Imputering av data betyr å sette inn estimater for manglende verdier i datasettet. Disse verdiene estimeres basert på de andre variablene i analysemodellen. Dette gjøres vanligvis med lineære regresjonsmodeller for kontinuerlige variable, f.eks. vekt, og med logistiske regresjonsmodeller for dikotome variabler, f.eks. samboerstatus. Disse regresjonsmodellene samlet kalles for *imputeringsmodellen*. Man lager flere komplette datasett, der det tas hensyn til at de imputerte verdiene er usikre, slik at de varierer mellom datasettene. Generelt anbefales det at man lager mellom 20 og 100 komplette datasett (2). Antallet avhenger sterkt av omfanget av manglende data. For å være på den sikre siden kan man gjerne bruke 100 imputerte datasett, med mindre det medfører at beregningstiden blir for lang.

Analysemodellen

Etter imputering analyseres hvert av de komplette datasettene med analysemodellen. Hver analyse gir et estimat med tilhørende varians for størrelsen(e) vi er interessert i, f.eks. regresjonskoeffisienten for fysisk aktivitet. Til slutt kombineres estimatene og variansene ved hjelp av spesifikke regler, de såkalte Rubins regler (1). Det kombinerte estimatet er lik gjennomsnittet av alle estimatene. Den kombinerte variansen er lik gjennomsnittet av variansene, pluss et ledd som tar hensyn til variasjon mellom de imputerte datasettene. Deretter kan man beregne konfidensintervall og *p*-verdi. Fremgangsmåten er illustrert i figur 1. Mange størrelser kan kombineres med Rubins regler, blant annet gjennomsnitt, standardavvik, andel og regresjonskoeffisient. Oddsratio og hasardratio bør logaritmetransformeres før de kombineres.

Oppdiktede data?

Imputeringsmodellen må være grundig gjennomtenkt. Alle variablene som skal være med i analysemodellen, inkludert den avhengige variabelen, må være med i imputeringsmodellen. I tillegg kan man ta med tilleggsvariabler som kan kalles

hjelpevariabler, dersom man har tilgjengelig variabler som ikke er med i analysemodellen, men som er assosiert med variabler der man mangler data. Dersom analysemodellen inneholder ikke-lineære funksjoner eller interaksjoner, må dette håndteres særskilt (3). Dersom interaksjonen inneholder en dikotom kovariat, f.eks. kjønn, kan man gjøre imputeringen separat i en delfil for hvert kjønn, og deretter slå sammen de to imputerte filene (2). Alt i alt krever imputeringsmodellen vesentlig mer tenkning og beregningstid enn selve analysemodellen.

Innebærer multippel imputering at man dikter opp data som ikke finnes? Tvert imot. Alle deltakerne inngår i analysen, også de som mangler data på en eller flere av variablene. Og med multippel imputering oppnår man høyere statistisk styrke og mindre skjevhet i estimatene enn med en analyse begrenset til deltakere med komplette data.

REFERENCES

1. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011; 30: 377–99. [PubMed][CrossRef]

2. van Buuren S. Flexible imputation of missing data. 2 utg. Boca Raton, FL: CRC Press, 2018: 175-6.

3. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Med Res Methodol 2012; 12: 46. [PubMed] [CrossRef]

Publisert: 21. januar 2022. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.21.0772