Tidsskriftet
DEN NORSKE LEGEFORENING

Korrelasjon og regresjon

KOMMENTAR

ARNE HØISETH
E-post: arnhois@online.no
Arne Høiseth er konsulent.
Ingen oppgitte interessekonflikter.

Are Hugo Pripp har sannsynligvis ønsket å gi en lettforståelig fremstilling av korrelasjonsanalyser (1). Men, «det er ingen kongelig vei til geometri». Snarveier og tilnærmet korrekte fremstillinger kan bidra til å underbygge misforståelser, feilaktigheter og manipulering av forskningsbudskap.

Korrelasjon (R) og regresjon henger sammen og kan ikke skilles. Det er for upresist å si at de røde linjene i figur 1 (1) er tilpassede linjer; det er regresjonslinjer hvor summen av de vertikale avstander mellom x-verdiene og linjen, benevnt residualer, er minst. Beregning av R begynner med å summere kvadrerte residualer, (KR).

Forskjellen i de to plottene (1) er neppe forskjell i stigning på regresjonslinjen, men kun forskjell i x-verdienes enheter. Man kan aldri sammenligne regresjonslinjens stigning hvis man ikke har benyttet samme enheter og skala, eventuelt må en utføre standardisert regresjon.

R-verdien viser ikke avstanden mellom punktene og regresjonslinjen; slike avstander er en måte å vurdere overensstemmelse på. R-verdien benyttes ofte, helt feilaktig, til å gi uttrykk for god eller dårlig overensstemmelse. For å bergene R-verdien trenger vi også summen av kvadratene av forskjellene mellom x-verdiene og gjennomsnittet av x-verdiene, benevnelsen «kvadrert total» kan benyttes (KT). R-verdien er proporsjonal til KR/KT. Jo større «kvadrert total» (KT), desto bedre R-verdi og tilsynelatende bedre overensstemmelse. «Kvadrert total» (KT) øker alltid med økt bredde (range) på y-verdiene og følgelig vil R-verdien alltid øke med økt bredde på utvalget av objekter. Å sørge for stor bredde på utvalget av objekter er et vekjent knep for å få en god R og tilsynelatende god overensstemmelse. Overensstemmelse er upåvirket av utvalgsbredden.

Ved R=0,09 forklarer variasjon i y så lite som 0,0081 prosent av variasjonen i x. For å bergene forklaringsprosenten benyttes ikke R, men r2, altså: r2 = 0,092 = 0,0081.

Påstanden om at man kan benytte regresjon til å predikere det ene tallet fra det andre er for upresist. Man kan lage «predikerte x-verdier» fra y-verdiene, men ikke y fra x. Hva slike predikerte verdier kan brukes til er for meg uklart. Hvis det ikke er en lineær assosiasjon er ikke R nødvendigvis 0. En kurvet sammenheng kan gi en høy R.

Pripp viser at p-verdien henger sammen med antall observasjoner, men om en korrelasjon/assosiasjon er (medisinsk faglig) relevant kan uansett ikke vurderes basert på p-verdier.

Jeg utfordrer Pripp til å diskutere betydningen av at den uavhengige variabel (y) aldri er 100 % presis, noe korrelasjon og regresjonsanalysen forutsetter. Flere forhold nevnt over er forøvrig beskrevet i Lægeforeningens Tidsskrift i 1990. (2).

LITTERATUR:

1. Pripp AH. Pearsons eller Spearmans korrelasjonskoeffisienter. Tidsskr Nor Legeforen 2018.doi: 10.4045/tidsskr.18.0042. [CrossRef]

2. Høiseth A. Er statistiske analyser egnet ved vurdering av målinger?Er statistiske analyser egnet ved vurdering av målinger? Tidsskr Nor Lægeforen 1990; 110: 1968 - 71. [PubMed]

Publisert: 12. juni 2018. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.18.0445