OPEN Odense Patient data Explorative Network

Guide

til

Reduktion af filstørrelse

(optimering af kørselstid i Stata)

OPEN Odense Patient data Explorative Network

Når man arbejder med registerdata på Sundhedsdatastyrelsens Forskermaskine eller Danmarks Statistik, krypteres cpr numre og andre identificerbare variabler (fx idnr / randomiseringsnumre) for at sikre, at data ikke er direkte personhenførbare. Derudover har nogle registre en form for unik id, som bruges til at koble forskellige tabeller. I Landspatientregisteret bruges variablen "recnum" til at koble administrationstabellen (LPRADM) med undertabeller som fx diagnosekoder (LPRDIAG).

Et krypteret cpr nummer har får en længde på 32 tegn på forskermaskinen ved SDS og 12 tegn på Danmarks Statistiks forskermaskine. Recnum fylder 12 tegn på Forskermaskinen ved SDS, mens det fylder 20 tegn på Danmarks Statistiks forskermaskine.

Ulempen ved at nogle variabler er meget lange er dels, at det kan tage lang tid at køre en enkelt kommando, og dels at datasættet optager meget plads, hvilket kan betyde, at man kommer til at opleve pladsmangel på den server, man arbejder på. OBS, det er muligt at tilkøbe 200 gb ekstra plads på Forskermaskinen

Formålet *med denne "Guide" er at give dig en hjælpende hånd til, hvordan du kan reducere længden på en variabel og dermed reducere størrelsen på den fil, du arbejder med.*

Derudover gives en kommando til, hvordan du kan arbejde med et lille udsnit af dit datasæt i dataklargøring- og analyseprocessen – med det formål at optimere kørselstiden.

Guiden er skrevet med udgangspunkt i, at statistikprogrammet Stata benyttes. Til denne guide er knyttet en Stata-do-fil *"Reduktion_filstørrelse"*. Elementer i denne do-fil gennemgås herunder.

Send en mail til OPEN.registry@rsyd.dk, hvis du ønsker, at vi skal sende dig denne do-fil
****

Do-filen er baseret på, at der anvendes data fra cpr registeret og andre nationale registre, men kan også anvendes til egne data.

Her er der nogle overvejelser, som du skal tage stilling til inden du begynder:

- Hvis du har et datasæt, der består af unikke cpr numre, kan du med fordel generere en ny variabel (id), der giver et fortløbende nummer til hver række.
- Hvis du har et datasæt, hvor der er dubletter eller hvor du ikke er interesseret i at lave en ny variabel, kan du anvende en metode, hvor du forkorter længden på en variabel til den mindst mulige, hvor den enkelte værdi stadig er unik. Denne metode fungerer bedst, hvis du vil reducerer en lang variabel som fx krypterede cpr numre eller variablen "recnum". Jo færre personer der indgår i datasættet, jo mere kan du reducere lænden på variablen. Derimod kan metoden have sine begrænsninger i forhold til originale cpr numre, hvor numrene, hvis der indgår mange personer i datasættet, kan risikere først at være unikt på et af de sidste cifre i cpr nummeret.

God arbejdslyst!

Har du forslag til forbedringer af denne guide, tager vi med glæde imod forslag. Kontakt os meget gerne på (OPEN.registry@rsyd.dk). OPEN fralægger sig ansvaret for eventuelle fej i do-filen.