Styregruppen for Højkvalitetsdata

23. juli 2008

Dokumentationsvejledning

Højkvalitetsdata: Dokumentation, videndeling mv.

Styregruppen for højkvalitetsdata består af: Hans Hummelgaard (fmd.) (akf og medlem af KOR), Helena Skytt Nielsen (Institut for Økonomi, AU), Flemming Petersson (Danmarks Statistik) og Charlotte Leolnar Reif (Danmarks Statistik).

1. Indledning

Målet med KOR og Danmarks Statistiks dokumentation af højkvalitetsdata er, at data, algoritmer mv. skal være veldokumenteret over tid. God dokumentation er med til at sikre kvaliteten af forskningen, idet risikoen for fejltolkninger mindskes, ligesom en god dokumentation også letter forskningsarbejdet. Ingen er interesseret i at offentliggøre resultater og komme med politikforslag, der bygger på misforståede data.

Målet med dette notat er at specificere, hvad der skal være de centrale elementer i dokumentationen. Der skal bl.a. være tabeller med randfordelinger og grafer, der viser den tidsmæssige udvikling i den respektive variabel. Endvidere skal der være en redegørelse for de fejlretninger, beregninger mv., der er sket siden dataindsamlingen. Der vil så vidt muligt blive registreret specielle forhold vedrørende fx administrationen af en given ydelse, der kan have betydning for tolkningen af tallene.

De respektive højkvalitetsdata vil i hvert fald indtil videre blive lagret i en SAS-databank, mens dokumentationen vil findes i Danmarks Statistiks generelle dokumentationssystem, TIMES, der vil blive udvidet til at kunne håndtere dokumentation tilbage i tiden. TIMES vil være tilgængelige for alle via internettet, og der er en særlig indgangsport til højkvalitetsdata.

2. Hvordan den enkelte variabel skal dokumenteres

Helt overordnet skal dokumentationen give forskerne adgang til den information, der er nødvendig for, at forskeren kan bruge oplysningerne korrekt. Al relevant information om variablen skal således være tilgængelig i en let forståelig form, let at finde og være overskuelig. Dokumentationen skal være tilgængelig i elektronisk form og indpakket i en grænsebrugerflade, der gør det let at finde de oplysninger, der er behov for, på en overskuelig måde. Det er naturligt at en del af dokumentationen bygger videre på allerede eksisterende dokumentation i Danmarks Statistik, men det er vigtigt at dokumentationen i så fald tilrettes, således at den er forståelig for målgruppen af forskere.

Dokumentationens indhold

Dokumentationen vil komme til at bestå af følgende elementer:

- Kort beskrivelse/label
- Udførlig forklaring
    - Hvad indeholder variablen, og hvordan er den konstrueret?
    - Hvilke personer variablen er observeret for i forhold til det grundregister, variablen stammer fra
    - Hvornår variablen er missing
    - Hvis variablen er konstrueret ud fra andre variabler, vil der være henvisninger/adgang til en beskrivelse af disse i det omfang, det er vigtigt for dokumentationen af den pågældende variabel
    - Hvis definitionen har ændret sig over tid, skal det fremgå af dokumentationen
- Beskrivelse af variablens udfaldsrum/værdisæt
- Statistisk beskrivelse af variablen for hele dataperioden, i form af tabeller med randfordelinger, middelværdier og fraktiler mv. ligesom der vil være grafiske præsentationer til at give et hurtigt overblik. For at lette overblikket i tabeller med

mange rækker indsættes subtotaler, og det oplyses hvilke grupper, disse totaler består af. Det er vigtigt, at den statistiske beskrivelse integreres i dokumentationen således, at årsagen til udsædvanlige udsving i en variabels værdier over tid forklares grundigt.

Uddybning om de enkelte punkter

I den udførlige beskrivende tekst vil der blive redegjort for, hvordan variablen præcist er defineret. Hvis variablen bygger på andre variabler, skal navnene på de pågældende variabler anføres, og det er et ønske fra styregruppen, at der skal være direkte link til dokumentation for den pågældende variabel. Derudover er det meget vigtigt, at der nøje redegøres for, hvordan definitionen af variablen eventuelt er ændret over tid. Hvis der har været databrud, eller variablens udfaldsrum ændres over tid, skal det omhyggelig beskrives. Hvis der er databrud i en given variabel, skal der henvises til en udgave af variablen, hvor der ikke er databrud (hvis en sådan findes), og det er et ønske fra styregruppen, at der skal kunne klikkes direkte hen til den pågældende variabel.

Ved databrud "på tværs af variabler" skal der være direkte henvisninger til dokumentationen for de forskellige variabler, som en given variabel med databrud er opdelt i. Fx skal der for BESKST være henvisning til BESKST02 og omvendt. Der skal endvidere for sådanne variabler være tabelleringer og grafer for de undervariabler, som en given variabel er opdelt i ved databrud (i eksemplet skal der således være graf og tabel for henholdsvis BESKST og BESKST02).

Der skal være en beskrivelse af populationen for variablen og udfaldsrum/værdisæt. Det er centralt, at oplysninger herom har et klart informationsindhold. Der etableres et fælles SAS-formatbibliotek, hvorved hver forsker ikke selv behøver at indkode værdisættet for hver enkelt variabel. En ekstra fordel ved at etablere et fælles formatbibliotek er, at det kan vedligeholdes centralt. Ændres udfaldsrummet for en variabel, er det således kun nødvendig at rette dette ét sted.

En meget væsentlig del af dokumentationen er, at forskerne får mulighed for at se randfordelinger for de enkelte variabler over tid. Herved kan databrud og/eller ændringer i udfaldsrummet meget hurtigt identificeres. Disse tabeller vil derfor være et meget nyttigt supplement til den skrevne dokumentation og være med til at sikre, at fejl og mangler i data bliver opdaget meget hurtigere, jf. tabel 1 for et eksempel på en randfordeling for civilstand. Når der er tale om stikprøver (som i tabel 1), vil det endvidere blive suppleret med en grafisk fremstilling af randfordelingen inkl. konfidenceintervaller. Grafiske illustrationer vil alene blive lavet for kontinuerte variabler.

Tabel 1
Randfordelingen for variablen CIVST (civilstand), personer over 14 år, 10% af befolkningen

	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993
Enke	41312	43237	45070	46830	48559	49963	51233	52547	53776	54969
Fraskilt	28508	29592	30702	31596	32414	33341	34305	34960	35530	35967
Gift	224143	222670	221147	220034	218982	218121	217253	216674	215408	214339
Længstlevende af 2 partnere	.	.	.	.	.	.	.	.	9	17

Ophævet partnerskab	.	.	.	.	.	.	.	.	4	8
Registreret partnerskab	.	.	.	.	.	.	.	6	189	213
Ugift	130105	132771	135493	137999	140271	142007	143920	146342	146948	148091
Alle	424068	428270	432412	436459	440226	443432	446711	450529	451864	453604

For variabler som fx uddannelse, branche mv. vil randfordelingen blive udskrevet på et aggregeret niveau (noget i retning af 10-15 undergrupper), da udskrivning på det mest disaggregerede niveau vil blive helt uoverskueligt. For variabler, hvor der er mulighed for forskellige aggregeringsniveauer (fx uddannelse, branche og DISCO) vil der blive redegjort for det i dokumentationen.

Hvis der i grafer eller tabeller er usædvanlige udsving, er det meget vigtigt, at der i dokumentationen redegøres for, hvad der er årsag til dette. Er det ændringer i lovgivningen, administrationen, definitionen af den pågældende variabel mv. Lovændringer mv. skal omtales i det omfang, at det er vigtigt for forståelsen af udviklingen i værdisættet for variablerne og/eller tolkningen heraf. Det er således særdeles centralt for kvaliteten af den registerbaserede forskning, at forskerne kender årsagerne til usædvanlige udsving i variablerne. Tabeller og grafer skal således integreres i dokumentationen.

Dokumentationen vil blive tilgængelig i elektronisk form, således at den er nem at overskue og giver et godt overblik over de data, der er defineret som HKD. Dette gøres ved at udvikle en overskuelig og brugervenlig grænsebrugerflade, hvor brugerne får adgang til forskellige typer oplysninger ved at klikke sig frem. Grænsefladen vil endvidere blive anvendt til at overskue data bedre ved at gruppere variabler efter emne og i alfabetisk orden. Variablerne grupperes efter følgende overordnede emner:

- Demografik
- Uddannelse
- Arbejdssteder og ansættelser (IDA)
- Arbejdsmarkedet
- Indkomster og løn
- Indkomstoverførsler
- Firmastatistik
- Boligstatistik

Dokumentationens form

Teksten i alle dele af dokumentationen skal skrives så forståelig, at også brugere uden forhåndskendskab til den pågældende variabel skal kunne læse teksten uden problemer. Det betyder fx, at fagspecifikke begreber skal defineres i teksten (eller i noter).

Der skal være en central læservejledning til alle felter i dokumentationsskabelonen

Al væsentlig information skal være i dokumentationen. Er uddybende oplysninger om forhold for forståelsen af variablen meget omfangsrig, kan der linkes til en uddybende

tekst herom fx en statistisk efterretning eller en uddybende tekst, der er udarbejdet i forbindelse med dokumentationen af den pågældende variable.

3. Videndeling

Styregruppen lægger stor vægt på, at der bliver gode muligheder for, at forskerne kan dele deres viden, erfaringer mv. om de enkelte variabler med hinanden. Videndelingen skal sikre: a) at de 'lokale eksperter' bliver gjort til 'globale eksperter', b) at forskerne kan indsende kommentarer og erfaringer med variablene. Konkret skal der være mulighed for, at forskerne kan skrive deres kommentarer mv. til en given variabel.

Danmarks Statistik stiller et videndelingprogram til rådighed, der kan køre parallelt med TIMES således, at det er muligt let at "svitse" fra det ene program til det andet. I videndelingprogrammet skal der meget let kunne søges efter alle kommentarerne for en given variabel, ligesom programmet automatisk skal indsætte navn, institution, dato for kommentaren og email-adresse for den forsker, der har indlagt kommentaren for en given variabel (det skal således være muligt for de enkelte forskere ved anvendelse af password selv at indlægge kommentarer i programmet). I forbindelse med dokumentationen af de respektive variabler vil der være en oplysning, om der findes forskerkommentarer til variablen. Er det tilfældet, og der klikkes på linket til videndelingprogrammet, skal brugeren komme direkte til kommentarerne for variablen (samtlige kommentarer til den respektive variabel vil være samlet ét sted).

Styregruppen fastlægger en procedure for, at kommentarerne bliver gennemgået med jævne mellemrum med henblik på at tage stilling til, om nogle af pointerne i kommentarerne skal medtages i den officielle dokumentation. I så fald placeres kommentarer i et bilag til dokumentationen.