Kursus i registerforskning

Kenneth Lykke Sørensen[1]

[1]Institut for Økonomi, Aarhus Universitet

Dagen i dag

Kursusbeskrivelsen dækker læringsmålene for dagens forløb.

09:00 - 09:30	Kaffe og rundstykker
09:30 - 09:50	Introduktion til KOR
09:50 - 12:00	Oplæg og øvelser - I
12:00 - 12:45	Frokost
12:45 - 14:00	Oplæg og øvelser - II
14:00 - 14:30	Kaffe og kage
14:30 - 15:45	Oplæg og øvelser - III
15:45 - 16:00	Opsamling

Agenda

1. Introduktion
2. Registerdata
3. Persondataloven
4. Øvelser
5. Den nye Persondataforordning fra EU (GDPR)
6. Sikkerhedsprocedurer
7. Øvelser
8. Adgang til data
9. At arbejde på en forskermaskine
10. Projektdatabaser

1. Introduktion

Introduktion - Hvorfor et kursus om registerdata?

- Demystificere registerdata
  - Registerdata er et stort område, hvor det kan være svært at finde rundt.
  - Men unik mulighed for at koble data fra forskellige områder.
- Overholdelse af datasikkerhedsregler
  - Tillid fra befolkningen
  - Uheldige sager - vi rammes alle når de brydes
- Den nye Persondataforordning fra EU (GDPR)
  - 'Ejerskabet' af persondata ligger nu entydigt hos personen selv

2. Registerdata

Kort historie om registerdata

Opbygningen af en internationalt set unik registerdatakultur i Danmark startede i slutningen af 70'erne og er siden blevet udvidet betragteligt.

Især op gennem 80'erne gik registerdata i Danmark fra at være en mindre sample af den danske befolkning, til med forskningsservice's oprettelse i 1988 at blive en konstant udvidende masse.

Se bl.a. festskriftet for FSEs 25 års jubilæum.

Leverandører af registerdata

1. Danmarks Statistik (DST)
2. Sundhedsdatastyrelsen (SDS)
3. Styrelsen for IT og Læring (STIL - Uddannelse oplysninger)
4. Andre styrelser (f.eks. STAR)
5. ...

Typer af registre

DSTs registeroversigt

1. Individdata (demografi, indkomst, uddannelse)
2. Virksomhedsdata
3. Arbejdsmarked
4. Uddannelse
5. Sundhedsdata
6. Andet

Dokumentation af registre

- DST
  - Højkvalitets dokumentation
  - TIMES
  - TIMES historiske data
- SDS
  - SDSs registre dokumentation på eSundhed.dk
  - LPR, LMDB, Sygesikringsregistret, osv.
  - SKS-browser

Data på individdata

1. Demografi (BEF, FAIN, FAFA)
   - Typer af familie: (C,D,E)
   - Antal af børn
   - Civil status

2. Indkomstregistret (IND). Detaljeret oplysninger på årbasis om indkomst på detaljeret og aggregeret niveau.
   - På familieniveau (D,E)
   - Per type
   - Skattebetaling og overførselsindkomst

3. Uddannelse (UDDA/BUE). Oplysninger om uddannelse ultimo året (altså oktober).
   - Højest fuldførte og igangværende uddannelse
   - Karakterer, tidspunkt for slut, uddannelsens normerede længde

Virksomhedsdata

- Generel firmastatistik
- IDAS, FIDA
- Regnskabsstatistik
- UHDI/UHDM/VARS/VARK

Arbejdsmarked

1. DREAM
   - DREAM er en sammenkobling af forskellige registre og oplyser for hver uge modtagelse af overførselsindkomster. Registret går tilbage til 1991 (uge 32). Den er lavet af beskæftigelsesministeriet og styrelsen for arbejdsmarked og rekruttering (STAR). Selvom registret ikke stammer fra DST kan den leveres af forskningsservice. Alternativt til undersøgelse af sygefravær kan man bruge RSS, som er lavet af NFA.
2. RAS - Registerbaseret arbejdsstyrkestatistik.
3. e-Indkomst/BFL/ILME. Oplysninger på månedsbasis om løn (BFL) og overførselsindkomster (ILME). Kan bruges til undersøgelser på beskæftigelsesområdet og Cost-benefit analyser.
4. Den Integrerede Database for Arbejdsmarkedsforskning (IDA): Erhvervserfaring, stillingskategorier, beskæftigelseskode, fagkode (discokoder), branche, osv.
5. Offentligt forsørgede (OF).

Arbejdsmarked: IDA

Uddannelse

1. Komprimerede elevregister (KOET, KOTO, KOTRE)
2. Grundskole
3. Folkeskolekarakter (UDFK)
4. Institutionsregister (INST)

Sundhedsdata

Vi nævner her de tre mest anvendte. Disse registre bliver leveret af DST og SDS

1. Landspatientregistret (LPR + LPRPSYK)
2. Lægemiddeldatabasen (LMDB)
3. Sygesikringsregistret (SYSI/SSSY)

LPRs og LMDBs dokumentation findes på eSundhed.dk

Landspatientregistret - LPR

LPR består af forskellige tabeller. Dokumentationen findes på eSundhed.dk. LPR går tilbage til 1977.

1. Administrative oplysninger (POP)
2. Diagnoser (DIAG siden 1994 ICD-10 koder)
3. Operationer (OPR)
4. Undersøgelser og behandling (UBE)
5. DAGS, DRG: afregning af stationære og ambulante behandlinger (går tilbage til 1995). Spørg en sundhedsøkonom eller Sundhedsministeriet. Det er en videnskab i sig selv.

Lægemiddeldatabasen - LMDB

Dokumentation findes her og her

1. Registrerer køb af receptpligtig medicin.
2. Går tilbage til 1995.
3. LMDB er omfattet af en yderligere lov, nemlig apotekerloven.
4. Særlige krav omkring adgang til LMDB og kobling til de øvrige registre.
5. SDS giver tilladelse til at få adgang og sætter eventuelle betingelser (diskretionering af visse variabler som fødselsdatoen f.eks.).

LMDB: ATC koder og øvrige variabler

- ATC: Anatomisk terapeutisk kemisk klassifikation.
- ATC koder: 5 niveauer.
- 5. niveau: produkt og derfor den mest følsomme. Man bruger den typisk til undersøgelser af effekten af et bestemt produkt.
- Dato for køb, pris (ekspeditionspris og patientbetaling) og volume (antal pakker, døgndosis, osv.).

Sygesikringsregistret - SYSI/SSI

- Dokumentation Sygesikringsregistret.
- Administrativ og afregningsværktøj.
- Registrerer hver ydelse, der bliver udført af læger samt bruttohonorar forbundet med ydelsen.
- Ikke nogen diagnoseoplysninger.

Andet

- Dødsårsagsregistret
- Boligoplysninger (BOL, BBR)
- Kriminalitet
  1. Afgørelser
  2. Indsættelser
  3. Konfererede sager
  4. Ofre for straffelovsforbrydelser
- Og meget meget mere ...

Praktisk info

De praktiske aspekter af brug af registerdata findes på følgende adresser:

- DSTs forskningsservice http://www.dst.dk/da/TilSalg/Forskningsservice
- SDSs forskerservice https://sundhedsdatastyrelsen.dk/da/forskerservice

INSTITUT FOR ØKONOMI
AARHUS UNIVERSITET

3. Persondataloven

Persondataloven

Al arbejde med registre til brug i forskningsprojekter er underlagt Persondataloven.

Læg især mærke til

§1 Loven gælder for behandling af personoplysninger, som helt eller delvis foretages ved hjælp af elektronisk databehandling, og for ikke-elektronisk behandling af personoplysninger, der er eller vil blive indeholdt i et register.

§3.1 Personoplysninger: Enhver form for information om en identificeret eller identificerbar fysisk person (den registrerede).

§3.1 Register med personoplysninger (register): Enhver struktureret samling af personoplysninger, der er tilgængelige efter bestemte kriterier, hvad enten denne samling er placeret centralt, decentralt eller er fordelt på et funktionsbestemt eller geografisk grundlag.

§5.5 Indsamlede oplysninger må ikke opbevares på en måde, der giver mulighed for at identificere den registrerede i et længere tidsrum end det, der er nødvendigt af hensyn til de formål, hvortil oplysningerne behandles.

§6.7 Behandling af oplysninger må kun finde sted, hvis behandlingen er nødvendig for, at den dataansvarlige eller den tredjemand, til hvem oplysningerne videregives, kan forfølge en berettiget interesse og hensynet til den registrerede ikke overstiger denne interesse.

Persondataloven

§7.1 Der må ikke behandles oplysninger om racemæssig eller etnisk baggrund, politisk, religiøs eller filosofisk overbevisning, fagforeningsmæssige tilhørsforhold og oplysninger om helbredsmæssige og seksuelle forhold.

§8.1 For den offentlige forvaltning må der ikke behandles oplysninger om strafbare forhold, væsentlige sociale problemer og andre rent private forhold end de i §7, stk. 1, nævnte, medmindre det er nødvendigt for varetagelsen af myndighedens opgaver.

§10.1 Oplysninger som nævnt i §7, stk. 1, eller §8 må behandles, hvis dette alene sker med henblik på at udføre statistiske eller videnskabelige undersøgelser af væsentlig samfundsmæssig betydning, og hvis behandlingen er nødvendig for udførelsen af undersøgelserne.

§10.2 De af stk. 1 omfattede oplysninger må ikke senere behandles i andet end statistisk eller videnskabeligt øjemed. Det samme gælder behandling af andre oplysninger, som alene foretages i statistisk eller videnskabeligt øjemed, jf. §6.

Persondataloven

§35.1 Den registrerede kan til enhver tid over for den dataansvarlige gøre indsigelse mod, at oplysninger om vedkommende gøres til genstand for behandling.

§35.2 Hvis indsigelsen efter stk. 1 er berettiget, må behandlingen ikke længere omfatte de pågældende oplysninger.

§43 Forinden iværksættelse af en behandling af oplysninger, som foretages for den offentlige forvaltning, skal der af den dataansvarlige eller dennes repræsentant foretages anmeldelse til Datatilsynet, jf. dog §44. Den dataansvarlige kan bemyndige andre myndigheder eller private til at foretage anmeldelse på dennes vegne.

§45 Forinden behandling, som er omfattet af anmeldelsespligten i §43, iværksættes, skal Datatilsynets udtalelse indhentes, når

1. behandlingen omfatter oplysninger, der er omfattet af §7, stk. 1, og §8, stk. 1
2. behandlingen udelukkende finder sted med henblik på at føre retsinformationssystemer
3. behandlingen udelukkende finder sted i videnskabeligt eller statistisk øjemed eller
4. behandlingen omfatter sammenstilling eller samkøring af oplysninger i kontroløjemed.

Apotekerlovgivningen

LMDB er udover persondataloven også underlagt apotekerlovgivningen. Denne er i løbet af de seneste år blevet revideret således at det for forskere er blevet betragteligt nemmere at få adgang til at koble LMDB med andre registre samt generelt at få adgang til oplysninger herfra.

Generelt gælder det dog at det kun er forskere der falder i en eller flere af nedenstående kategorier, der kan opnå adgang til LMDB:

1. Personer der er ansat i det offentlige sundhedsvæsen
2. Personer der praktiserer efter overenskomst på sundhedsområdet
3. Forskere der er ansat på et universitet
4. Forskere der er ansat i en patientforening

Forskere som falder udenfor personkredsen defineret ovenfor, kan iflg. Sundhedsdatastyrelsen kun få adgang til fuldt anonymiserede data på projektet, så det ikke er muligt at identificere patienten.

Apotekerlovgivningen

For at kunne opretholde fuld anonymitet, vil det iflg. Sundhedsdatastyrelsen være nødvendigt med restriktive adgange til data, hvor der kun gives adgang til de mest nødvendige variable underlagt strenge grupperinger. Ved små populationer, hvor det er nemmere at identificere enkelte personer, vil der blive givet adgang til færre oplysninger end ved større populationer.

Efter aftale med Sundhedsdatastyrelsen ændres der ikke i afgørelser afgivet før den 1. november 2016. Dette gælder også for de tilknyttede dataansvarlige på projekterne, hvor tilladelsen er givet før 1. november 2016. Det er først hvis projekterne kommer i høring hos Sundhedsdatastyrelsen igen i forbindelse med udvidelser eller andet, at de nye regler vil træde i kraft for eksisterende projekter.

Etiske overvejelser

Det er i en række tilfælde (især når der indgår sundhedsdata) nødvendigt at søge om en etisk tilladelse til at gennemføre sit projekt.

Fra Den Nationale Videnskabsetiske Komité:

> *Et sundhedsvidenskabeligt forskningsprojekt, der involverer mennesker eller menneskeligt biologisk materiale så som væv, æg og celler skal godkendes af en videnskabsetisk komité, inden forskningsprojektet sættes i gang.*

4. Øvelser

Øvelser

I grupper af 4-5 personer, diskuter for projekt (a), (b) og (c):

1. Hvilke (hvis nogen) etiske overvejelser ser I i forbindelse med projektet?
2. Hvilke typer registre kunne være relevante for jer at søge om adgang til i forhold til projektet?
3. Hvilke typer af resultater kunne I forestille jer I gerne vil hjemtage i forbindelse med projektet?
4. Hvad skal man særligt være opmærksom på under dette projekt?

(a) I ønsker at undersøge effekten af ekstra tandeftersyn til folkeskoleelever fra socialt udsatte hjem, hvor det interessante outcome er den fremtidige trivsel for den enkelte elev.

(b) I har indsamlet et spørgeskema foretaget blandt praktiserende læger og deres patienter, som I har fået lagt ind på et projekt, hvor det skal kobles med data fra Danmarks Statistik og SDS, så I kan undersøge om patienter aktivt vælger læge efter lægers tilbøjeligheder til at godkende sygefravær.

(c) I har et projekt der ser på effekten af kontanthjælpsreformen, og vil specifikt undersøge hvilke effekter den har haft på helbredet af udsatte kontanthjælpsmodtagere.

5. Den nye Persondataforordning fra EU (GDPR)

Den nye Persondataforordning fra EU (GDPR)

Den nye Persondataforordning, også kaldet General Data Protection Regulation eller Databeskyttelsesforordningen træder i kraft 25. maj 2018. Den gælder data om samtlige fysiske personer (dvs. individer og enkeltmandsvirksomheder).

GDPR indeholder i vid udstrækning de samme regler, som vi allerede har i den danske Persondatalov.

Formålet med GDPR kan grundlæggende koges ned til:

- Sikre fri udveksling af personoplysninger
- Øge den digitale tillid
- Beskytte grundlæggende rettigheder

Den nye Persondataforordning fra EU (GDPR)

Konsekvenserne for forskere (og specielt forskernes arbejdsgivere) er især:

- Ansvaret når der behandles personoplysninger gøres større og mere gennemsigtigt
- Personerne bag personoplysningerne får betragteligt flere rettigheder (de 'ejer' reelt data om dem selv - dog ikke i de nationale registre)
- Datatilsynet har fået skærpede håndhævelsesmuligheder overfor databehandlere
- Straf og bøder i forbindelse med overtrædelse af reglerne øges markant - op til 4% af den årlige driftsbevilling, maks 16 millioner kr.

Den nye Persondataforordning fra EU (GDPR)

Der skelnes i GDPR (ligesom i Persondataloven) mellem

- Den dataansvarlige
  - En fysisk person, virksomhed, myndighed eller offentlig institution, der ønsker at få behandlet personoplysninger til bestemte formål.
  - Er ansvarlig for at reglerne bliver overholdt.
- Databehandleren
  - En virksomhed, offentlig myndighed eller fysisk person, der *ikke* er ansat hos den dataansvarlige, men som behandler personoplysninger på den dataansvarliges vegne og efter instruks fra den dataansvarlige.

Det betyder, at der altid skal foreligge en databehandleraftale mellem den dataansvarlige og databehandleren, hvis behandlingen af personoplysninger ikke foretages af den dataansvarlige selv. Dette er for eksempel tilfældet, når forskere opnår adgang til data via DST eller SDS, hvor universitetet, hospitalet, forskningscentret, eller lign., hvor forskeren er ansat er den dataansvarlige mens DST eller SDS er databehandler.

Den nye Persondataforordning fra EU (GDPR)

GDPR giver en række rettigheder til personen bag personoplysningen:

- retten til at få besked om, at der behandles personoplysninger. Det kaldes også oplysningspligt
- retten til at se sine oplysninger, hvilket kaldes indsigtsret
- retten til at få berigtiget forkerte personoplysninger, også kaldet retten til berigtigelse
- retten til at få begrænset behandlingen af sine oplysninger
- retten til at gøre indsigelse mod behandling af sine oplysninger, også kaldet indsigelsesretten
- retten til at blive slettet
- retten til dataportabilitet
- retten til ikke at være genstand for en afgørelse, som alene er baseret på automatisk behandling

Den nye Persondataforordning fra EU (GDPR)

Praktiske IT problemstillinger

- Du må aldrig opbevare eller behandle fortrolige eller følsomme personoplysninger på din private computer eller andet privat udstyr. Hvis du arbejder med personoplysninger, skal du altid benytte den computer, du som medarbejder har fået udleveret af din arbejdsgiver.
- Du må som udgangspunkt ikke opbevare persondata på Cloudtjenester (fx OneDrive, Dropbox, Google drev etc.), medmindre din arbejdsgiver har et gyldigt aftalegrundlag med den pågældende udbyder (AU har fx ikke sådanne aftaler).
- Så længe du arbejder med personoplysninger, er det bedste sted at opbevare dem på din arbejdsgivers netværksdrev. På den måde sikrer du, at personoplysningerne opbevares forsvarligt, og at du har en backup, hvis uheldet er ude.
- Hvis du ikke har mulighed for at opbevare personoplysninger på et netværksdrev, men fx er nødt til midlertidigt at opbevare oplysningerne på din udleverede computer, **skal du sikre dig, at computeren er krypteret**.
  - Det samme gælder for tilfælde, hvor du gemmer personoplysninger på en USB-nøgle.
  - Bemærk endvidere, at der med GDPR også er et krav om at al aktivitet med personoplysninger bliver logget!

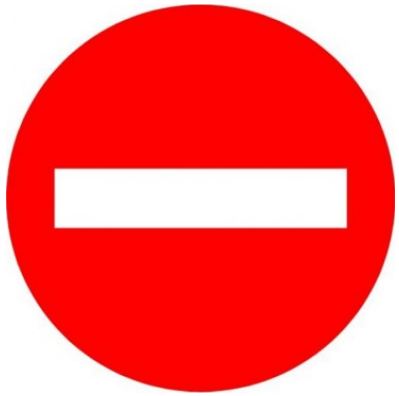
Den nye Persondataforordning fra EU (GDPR)

Generelt, så er GDPR langt hen ad vejen en bekræftelse af den danske persondatalov, men med skærpede rettigheder til personen bag personoplysningen samt med betragteligt skærpede sanktionsmuligheder overfor sikkerhedsbrud.

Endvidere pålægger GDPR skærpede krav på den dataansvarlige til at sikre at data behandles og opbevares IT-mæssigt forsvarligt.

6. Sikkerhedsprocedurer

Sikkerhedsprocedurer

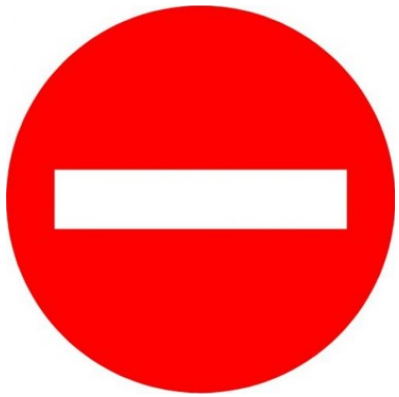


Hovedreglerne omkring hjemsendelse:

1. Hjemsendte resultater bør som udgangspunkt have sådan en form, at de umiddelbart kan anvendes i en publikation.
2. Man må ikke hjemsende mikrodata!
3. Det må ikke være muligt at identificere en person eller virksomhed i hjemsendt materiale
   - Minimum 3 observationer pr. celle i tabel eller punkt i graf
   - En eller to virksomheder må ikke udgøre 80+% af værdi i tabelcelle eller graf

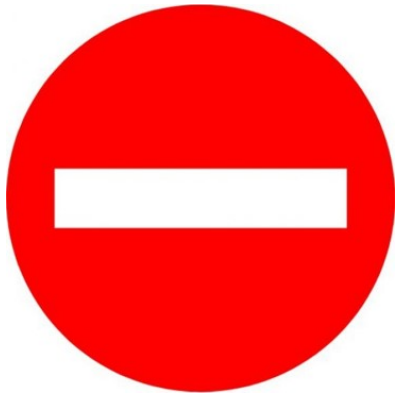
Se reglerne for hjemsendelse.

Generelle principper



1. Ikke tilladt at downloade individdata (bevidst eller ubevidst).
2. Forbudt at videregive sit password og adgang. 1-3 måneders karantæne i første omgang afhængig af om det er med vilje eller ej.
3. Ikke nogen logging af forespørgsler. Derfor er systemet baseret på tillid.
4. Indtil videre: "Menneskelig"stikprøvekontrol udført af DSTs forskningsservice
5. DST har et scanningsprogram, der tjekker filer der ønskes hjemsendt for 'usual suspects' - dette skal IKKE forstås som, at man så ikke selv behøver være opmærksom!

Generelle principper



Man må IKKE søge efter bestemte personer:

- Det kan være fristene og man synes sikkert, at der er en vældig god grund til at gøre det, MEN det er strengt forbudt!
- Lad være med at søge efter bestemte personer i registrene - heller ikke dig selv!

Hjemsendelse i praksis

- DSTs server:
  - Brug af et særligt program på serveren (forklaret her).
  - Overførslen sker med det samme.
  - Visse filtyper er ikke tilladt.

- SDSs server:
  - Udbakkesystem.
  - Send filerne i en mappe.
  - Hvert 10. minut bliver filerne sendt ude af serveren til brugerens emailadresse.

Må hjemsendes I

- Analyseresultater
- Aggregerede tabeller
- Figurer
- Det må ikke være muligt at identificere personer, husstande, familier, virksomheder og andre enheder selvom afidentificeret løbenummer.

Må hjemsendes II

- For aggregerede data: Hjemsendte resultater bør som udgangspunkt have sådan en form, at de umiddelbart kan anvendes i en publikation.
- Men det er tilladt at arbejde videre med det hjemsendte aggregerede materiale, fx til dannelse af figurer.
- For tabeller: Tabeller skal indeholde mindst 3 observationer pr. celle.
- Dette er en tommelfingerregel. Grænsen kan være højere i nogle tilfælde, hvis det fx er åbenlyst hvilke 3 personer der er tale om.
- Se følgende dokument DSTs datafortrolighedspolitik, kapitler 3 og 4 for flere detaljer.
- Diskretionering af rækker i praksis.
- og fjernelse af diskretionerede rækker i praksis.

Må hjemsendes III

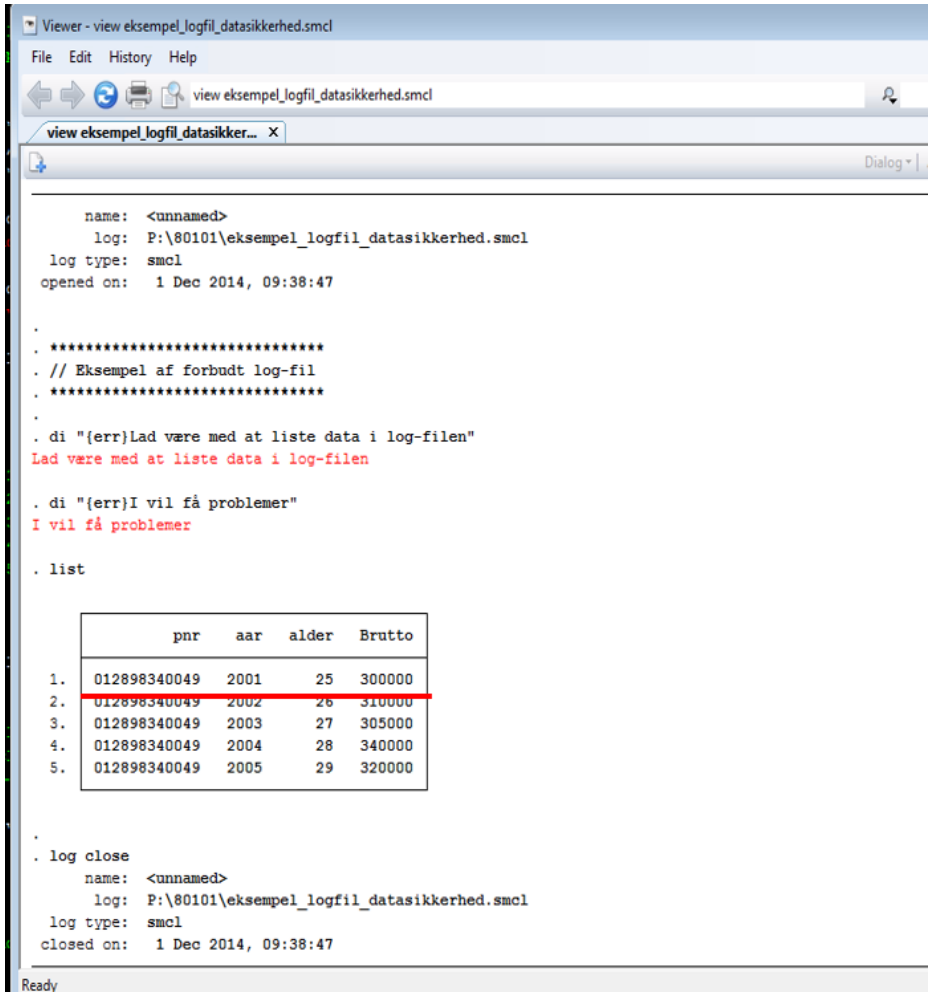
- Tabeller: Pas også på Max/Min-værdier og percentiler der i nogle tilfælde kan lede til identificerbar information (tænk for eksempel på indkomst).
- Figurer: Må kun hjemsendes, hvis de ikke indeholder identificerbar information. Tænk for eksempel på et scatterplot af indkomst, hvor ekstreme værdier er tydelige. I nogle tilfælde kan man identificere bestemte individer.
- Vær generelt opmærksom på ekstreme værdier.

Må IKKE hjemsendes I

- Al arbejde med mikrodata skal foregå på forskermaskinerne
- Enhver fil, der indeholder data på individniveau, selv hvis data er anonymiserede, må ikke hjemsendes
- Afidentificerede nøglevariabler som personnummer, CVR-numre, mv.
- Programkoder og logfiler, der indeholder individdata eller nøglevariabler
- Hvis du har fået koblet data ind på serveren du selv har indsamlet, så gælder de samme regler - dvs. du må heller ikke hjemsende data du selv har fået lagt derind.

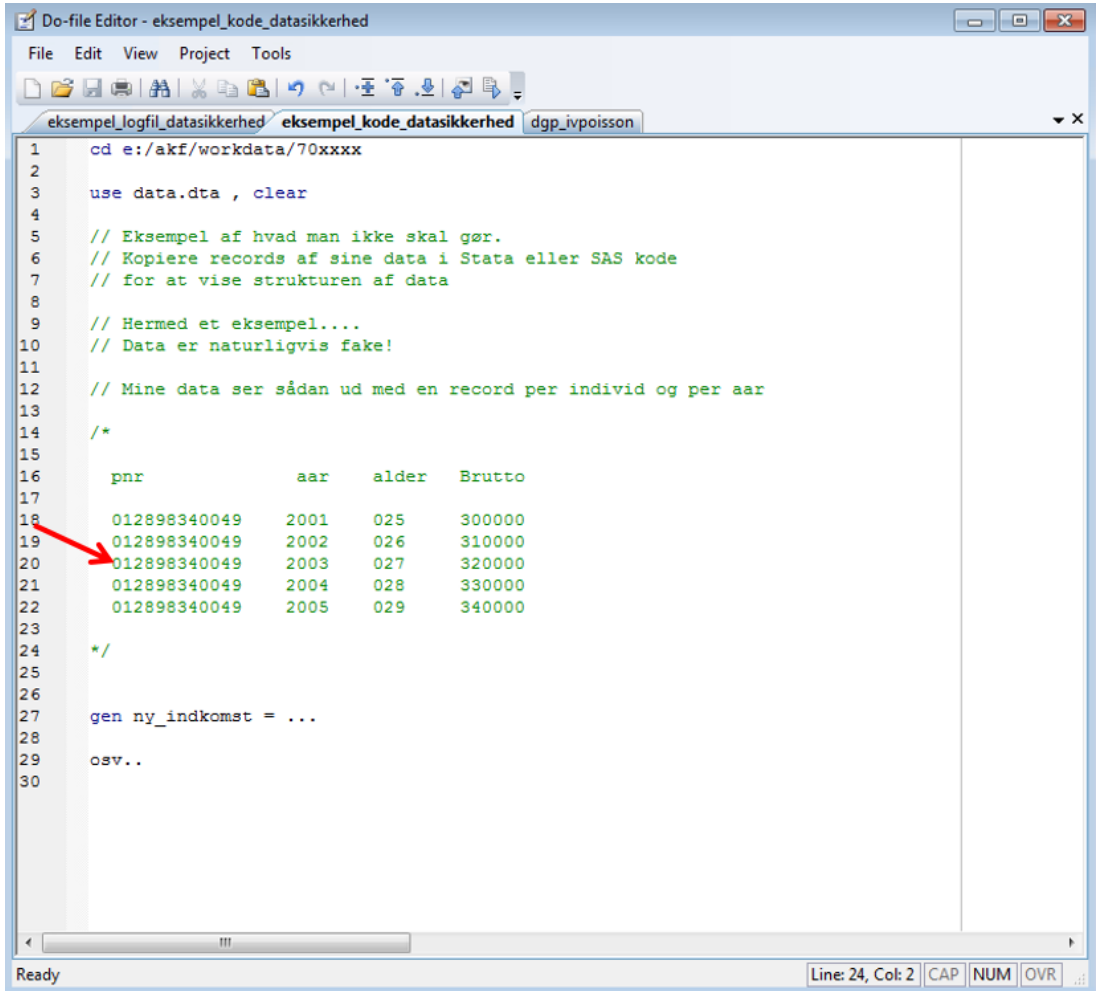
Må IKKE hjemsendes II

I må gerne lave Log-filer, der indeholder en listing af data MEN de må **ikke** hjemsendes.



Må IKKE hjemsendes III

Kode, der indeholder kopier af data må ikke hjemsendes



Sanktioner for enkelte medarbejdere, der overtræder reglerner

Danmarks Statistiks regler er fastsat i autorisationsaftalen og aftalen mellem DST og den enkelte registermedarbejder.

> **"En overtrædelse af reglerne i denne aftale vil medføre, at den pågældende forsker udelukkes fra at anvende nogen af Danmarks Statistiks forskerordninger permanent eller i en periode på ikke under 3 år. Endvidere vil en overtrædelse medføre, at nærværende autorisation bortfalder i en periode."**

(§12 i autorisationsaftalen)

Procedurer ved et regelbrud: Hjemsendelse af mikrodata

- Forskernes adgang til mikrodata inddrages for samtlige projekter, der hører under den institution, der **ejer** det konkrete projekt.
- Denne regel gælder også, selvom forskeren hører til en anden institution end den, der ejer projektet.
- Karantæne fra 1-3 måneder afhængig af alvorligheden af overtrædelsen af reglerne.
- Forskeren der overtræder reglerne risikerer permanent frakendelse af ret til adgang til registerdata betinget på alvoren af overtrædelsen.

Hvis der begås et regelbrud

- Orienter altid snarest Danmarks Statistik, din datamanager, din analyse-og-forskningschef og evt. din projektleder ved et regelbrud.
- Beskriv i mailen hvornår data er sendt hjem samt omfanget af hjemsendelsen. Adgangen til mikrodata vil fortsat blive lukket og der skal sendes en redegørelse og plan for forebyggelse.
- Orienteres der hurtigt om et regelbrud betragtes det som en formildende omstændighed.

Tjek af resultater

For at overholde sikkerhedsprocedurerne skal du altid sikre, at de resultater du sender hjem overholder datasikkerhedsreglerne.

Det er dit ansvar, og straffen for overtrædelse er som udgangspunkt den samme uanset om du overtræder reglerne med vilje eller på grund af uopmærksomhed.

Hvordan kan vi minimere risiko for fejl/brud på datasikkerhedsregler

Hermed et forslag til hvordan vi kan sikre, at vi kun sender autoriseret output hjem.

- Kopier de resultater/filer, som I ønsker at hjemsende til en mappe, f.eks ∼/toSend
- Tjek samtlige filer i mappen
- Send filerne enten med DSTs program eller ved at kopiere filerne i din udbakke (SDSs server).

Begræns jer til det mest nødvendige, dvs. prøv at bearbejde resultaterne mest muligt på serveren, så I sender en tabel, der kan indsattes i et dokument. SAS (ODS), Stata (estout/outreg) og R har programmer, der kan hjælpe i denne fase.

Gode råde til forskeren, der har adgang til mikrodata

- Kontroller altid samtlige filer, der skal hjemsendes
- Kontakt altid Forskningsservice i tvivlstilfælde
- Begræns hjemsendelsen af filer til det mest nødvendige.
  - Er det fx nødvendigt at hjemsende programmer og logfiler? De opbevares sikkert under forskerordningen
  - Er det nødvendigt at hjemsende foreløbige resultater?
- HUSK at analyser på mikrodata skal foretages på Danmarks Statistiks (eller SDSs) forskermaskiner, og må ikke hjemtages med henblik på videre analyser andetsteds.

7. Øvelser

Øvelser

Øvelser omkring hvilke filer der må hjemsendes udleveres.

Diskuter opgaverne i grupper af 4-5 personer.

Efter 20 minutter mødes vi og diskuterer dem i plenum.

INSTITUT FOR ØKONOMI
AARHUS UNIVERSITET

8. Adgang til data

Autorisering af institutter

Uddrag af DSTs hjemmeside:

"For at få adgang til mikrodata under Danmarks Statistiks forskerordning, skal det enkelte forsknings-/analysemiljø først autoriseres. Kun mere varige forsknings- og analysemiljøer med en ansvarlig chef og flere forskere/analytikere kan autoriseres. Enkeltmandsvirksomheder kan ikke autoriseres.

Ved hver autorisationsansøgning foretager Danmarks Statistik en konkret bedømmelse af miljøet. Her lægges særligt vægt på miljøets kompetencer indenfor håndtering af registerdata samt kendskab til de datasikkerhedsregler, der gælder for adgang til mikrodata under Danmarks Statistiks forskerordning."

Når forskeren søger adgang til data

Brugeroprettelse:
For at kunne opnå adgang til registerdata hos Danmarks Statistik skal man for det første være ansat af en autoriseret institution (f.eks. universitet, sektorforskningsinstitution eller firma, der bruger registerdata til statistiske analyser).

Dernæst indsendes der en forespørgsel om oprettelse af en ny bruger hos DST, der tilsender en forskeraftale til underskrift hos brugeren samt den ansvarlige på institutionen. Herefter tildeles forskeren en ident, og kan derefter tilskrives projekter som forskeren skal arbejde på.

Forskeraftalen

Uddrag af forskeraftalen

- "De datasæt, der gives adgang til, er fortrolige i henhold til Forvaltningslovens §27 [tavshedspligt], stk. 3 og Straffelovens §152 [uberettiget videregiver eller udnytter fortrolige oplysninger - strafferamme: bøde eller 1-6 mdr. fængsel]." (§1 i forskeraftalen)
- "...alt output skal være aggregeret i en sådan grad, at der ikke er fare for direkte eller indirekte identifikation af personer eller virksomheder. Der må ikke foretages forsøg på en sådan identifikation" (§7 i forskeraftalen)
- "De af Danmarks Statistik udleverede passwords er personlige og må ikke udlånes eller meddeles til andre" (§4 i forskeraftalen)

Adgang til mikrodata og Forskningsprojekter

Projektoprettelse:
Når man har en forskeraftale, og vil oprette et nyt projekt søges der om adgang til de data der er relevante for projektet.

- DST: procedurens beskrivelse og projektindstillingen findes her
- SDS: ansøgningsproceduren er forklaret her

Need-to-know princippet

Bemærk at ethvert forskningsprojekt er underlagt *Need-to-know* princippet hos Danmarks Statistik. Det vil sige, et projekt kan kun godkendes hvis der er gjort tilstrækkelig rede for nødvendigheden af hvert enkelt ansøgt register (og variabler).

Dette bunder i at de registre der kan søges om adgang til alle indeholder meget præcise oplysninger om enkeltindivider/virksomheder/institutioner osv. i Danmark, og i henhold til Persondatalovens §5 Stk. 3,

> *Oplysninger, som behandles, skal være relevante og tilstrækkelige og ikke omfatte mere, end hvad der kræves til opfyldelse af de formål, hvortil oplysningerne indsamles, og de formål, hvortil oplysningerne senere behandles.*

har en forsker ikke ret til at have adgang til andre oplysninger end de der er nødvendige for projektets udførelse.

Datatilsynet

For at sikre at Persondataloven overholdes, er det pålagt DST at sikre at alle relevante tilladelser er indhentet fra Datatilsynet inden et projekt oprettes.

Alle forskningsprojekter, hvor der behandles følsomme personoplysninger skal anmeldes til Datatilsynet.

Der er for et par år siden sket en ændring i hvordan der søges om tilladelse til at udføre et forskningsprojekt hos Datatilsynet, således at, hvis man er tilknyttet en større organisation (fx et universitet), så er behandlingen af anmeldelser til Datatilsynet nu blevet lagt hos organisationen.

Forskeraftaler

- Når man skifter arbejdsplads, så skal der skrives en ny forskeraftale - dette bunder i at den grundlæggende dataansvarlige er den personaleansvarlige på den autoriserede institution (på universitetet er det ofte institutlederen eller centerlederen).

- For samarbejdspartnere - der skal skrives en forskeraftale med alle der skal have adgang til et projekt.

9. At arbejde på en forskermaskine

DSTs server

Vejledninger og applet findes her :
http://www.dst.dk/da/TilSalg/Forskningsservice/Vejledninger

Udvalg af programmer er begrænset, ligesom hvad man kan gøre med maskinen. Adgang og skriverettigheder er stærkt begrænset. Derfor er brugen af serveren (lidt) anderledes end på en almindelige computer.

Der findes blandt andet følgende programmer:

1. Statistiske programmer: Stata, SAS, R, SPSS, WPS (en SAS-clone)
2. Utilities (StatTransfer, hjemsendelse af resultater)
3. Tekst editor (emacs med ESS, Sublime, Vim)

Hvert statistisk program har sin egen server. Derfor kan man ikke kalde et program, der ligger på en anden server. Man kan for eksempel ikke kalde Stata eller R ind fra et SAS program.

Struktur af et projekt

1. Hvert projekt får et projektnummer (70xxxx)
2. Når du logger ind, så gør du det på et specifikt projekt. Dvs. når du er logget ind, så kan du kun se det projekt du er logget ind på, selvom du måske har adgang til adskillige projekter.
3. Der bliver oprettet to mapper e:/70xxxx/Rawdata og e:/70xxxx/Workdata
4. Der er også et Z- drev. Den er lokal til projektet (dvs. man kan ikke få adgang til den fra et andet projekt). I må ikke lægge data i denne mappe. Det er en "privat"mappe (andre på projektet har ikke adgang). Undgå at lave analyser på denne mappe selv hvis I arbejder alene på projektet. Den er mere egnet til at have en særlig opsætning af de forskellige programmer (SAS-autoexec fil).
5. DST leverer data enten fra deres grundregistre eller fra en eventuel projektdatabase i Rawdata. Rawdata er skrivebeskyttet og I kan derfor ikke ændre denne mappes indhold.
6. Alt analysearbejde skal foregå i Workdata.

Hostede maskiner

- Mulighed for en institution for at have sin egen maskine
- Skal køre med Windows og skal fysisk placeres hos Danmarks Statistik
- DST er ansvarlig for vedligeholdelsen af serveren (men det er ikke gratis!)
- En ansat af den pågældende institution fungerer som administrator.
- Spørg forskningsservice eller DSTs IT-center om prisene.
- En del institutioner benytter sig af denne mulighed (SFI, IFSV, IfØ på Aarhus universitet, CAM fra IfØ på Københavns universitet).
- Datasikkerhedsreglerne er de samme som på FSE-serveren og logging på serveren tager samme udgangspunkt.

DSTs server: Logging på serveren:

DST har fået en ny metode til at logge på serverne, der ikke længere er afhængig af Java (mega godt!).

Vejledninger

- Gå til `https://remote.dst.dk`
- Indtast din ident uden projektnummer og din 4-cifrede pinkode (udleveres af DSTs forskningsservice når du underskriver din autorisationsaftale)
- Du bliver bedt om at indtaste et one-time password, som du modtager på din mobiltelefon (eller via token, hvis du er ældre bruger)
- Vælg den relevante maskine (FSE eller hostede)
- Du kan arbejde fra hjemmet uden en hjemmearbejdsplads (med en betingelse: du skal have en dansk ip-adresse)

SDSs server

SDS har også en forskermaskine

- Ordningen ligner meget DST's forskermaskine.
- Vejledninger
- Datasikkerhedsregler er grundlæggende de samme som på DST-servere.

Du er ikke alene

En forskermaskine kræver god opførsel af alle der arbejder på den

- Brug af ressourcerne: Vær opmærksom på dit dataforbrug (RAM og diskplads)
- Valg af algoritme kan afgøre performance og brug af ressourcer
- Vi deler ressourcerne, derfor påvirker og begrænser dit brug af ressourcer andres.
- Undgå at arbejde med eller minimer brugen af meget store datasæt.
- Undgå at køre flere (store) jobs på samme tid (eller kun hvis det er absolut nødvendigt).
- Nogle opgaver kræver flere ressourcer end andre. Det skal man være klar over.
- Sluk lyset om aftenen når du er færdig.
- Ryd op efter dig: Slet midlertidige filer (SAS og Stata).
- Vi deler licenser. Så undgå at åbne flere sessioner af samme program (dette er patologisk i Stata)

Råd omkring brug af ressourcer

- Sortering og transponering af data kræver mange ressourcer. At lære lidt programmering vil gøre dit liv nemmere og vil glæde andre, fordi du ikke bruger så mange ressourcer.
- Visse statistiske modeller er krævende i CPU tid (Proc glimmix i SAS f.eks) så overvej nøje, når I sætter en opgave i gang.
- Ryd op efter dig, når du er færdig
  - i SAS: tøm **Work**-biblioteket af udnøvendige datasæts, der fylder (proc datasets og delete statement)
  - i Stata: clear hukommelsen.

At arbejde med store datasæt

- At arbejde med store datasæt i Stata:
  - Large datasets
  - Stata big datasets NBER
  - Joe Canners Stata præsentation
- Brug af batch mode (kørsel i baggrund) med store jobs. Link til Stata
- Diskleje: diskplads har en omkostning
- Arbejd med mindre udgaver af dit datasæt.

Brug af DSTs formater

En del af DSTs formater ligger på forskermaskinen. Disse formater faciliterer brugen af variablerne fra DSTs grundregistre og viderekodning af disse variabler.
Vejledninger ligger på forskermaskinen (Stjerne-ikon) og på Forskningsservices hjemmeside.

- SAS: findes som SAS formater, SAS catalog, SAS datasæts
- Stata og SPSS: findes kun som datasæts. Kan flettes på datasættet.
- Sætter tekst på numeriske koder
- hjælper i at gruppere visse koder i mere aggregerede grupper
- f.eks. uddannelseskoder, branchekoder, oprindelsesland, region og kommuner, osv.

10. Projektdatabaser

Formål

DST giver mulighed for at oprette projektdatabaser

- Formål med en projektdatabase er for en institution at have fælles registerdata liggende på en server (FSE eller hostede maskiner).
- Denne database bruges til at lave udtræk til institutionens registerundersøgelser.
- Kræver en datamanager (mulighed for at have to datamanager).
- Stordriftsfordele ved at genbruge databasens fælles indhold og samle flere registre sammen.
- Mindsker behandlingstid i Forskningsservice.

Hvad laver en datamanager

- Vedligeholder databaser
- Laver dataudtræk til projekter
- Er kontaktperson mellem Forskningsservice og den pågældende institution

Hvordan fungerer det?

- Databasen er et selvstændigt *projekt*, som kun datamanageren har adgang til
- Datamanageren har to muligheder:
  - Han eller hun kan lave et fysisk udtræk af data med de relevante registre/variabler og for den relevante population
  - Eller lave det man kalder et SAS-View. Et SAS-view er en SAS-fil der definerer et datasæt. Det tillader at afgrænse datamaterialet til den rigtige population og de relevante variabler og fylder næsten ingenting. De sparrer også en masse tid, når det kommer til at lave udtræk.
- Variabeloversigten af Jeres eventuelle database vil ligge på e:/Institutions navn/Oversigter/.