Vejning mm. i R

Lolle, Henrik

Publication date:
2024

Document Version
Også kaldet Forlagets PDF

Introduktion til vejning mm. i 

Formidlingsartikel af Henrik Lauridsen Lolle
Institut for Politik og Samfund, Aalborg Universitet
Revideret i december 2024

Indhold

Indledning

I den mest basale statistiske teori om inferens fra stikprøve til population ligger der en forudsætning om *simpelt tilfældigt udtræk* af stikprøve bag, dvs. hvor stikprøvens enheder udvælges helt tilfældigt fra en liste over populationens enheder, og hvor alle populationens enheder har samme sandsynlighed for at blive udvalgt til stikprøven. Samme forudsætning ligger også bag default-versioner af statistiske analysemetoder i programmer som SPSS, SAS, Stata og R. Endvidere er det denne basale statistik og default analysekommandoer, der som oftest præsenteres i samfundsvidenskabelige metodebøger, som fx Alan Agrestis (2018) meget benyttede *Statistical Methods for the Social Sciences*. Meget ofte er survey-data imidlertid *ikke* udtrukket simpelt tilfældigt, og hvis de er, har der ofte været et skævt bortfald i form af manglende besvarelse af spørgeskemaet fra bestemte typer i populationen. På den måde er *netto-stikprøven*, også kaldet

analyserammen, ikke repræsentativ for populationen. Dette gælder også for survey-data fra store internationale survey-organisationer som ISSP, EWS, WVS og EES[1].

Heldigvis kan man alligevel, også med almindeligt brugte statistikprogrammer som de ovenfor nævnte, foretage statistisk inferens til populationen med en god portion ekstern validitet. Det forudsættes dog stadigvæk, at stikprøvens enheder er udvalgt med en eller anden form for *sandsynlighedsudvælgelse*, hvor man kender sandsynligheden for udvælgelse til stikprøven (brutto-stikprøven) for samtlige enheder i populationen. Desuden kan det i en del situationer være vigtigt, at man har forskellige andre typer af oplysninger om, hvordan stikprøveudtrækket er foretaget, herunder også karakteren af et evt. skævt bortfald fra brutto-stikprøve til netto-stikprøve.

I det følgende vil jeg give dels en kort introduktion til stikprøveudtræk fra en population, dels gennemgang af, hvordan man i programmet R ved brug af pakken `survey` (Lumley, 2023A) kan specificere sin stikprøve og efterfølgende foretage statistisk analyse på data, der ikke er baseret på simpel tilfældig udvælgelse af stikprøvens enheder, og hvor der evt. har været et skævt bortfald fra brutto- til netto-stikprøve. Analyseeksemplerne vil være nogle få univariate analyser samt regressionsanalyse og grafisk visning af kontrollerede effekter med pakken `marginaleffects` (Arel-Bundock, 2023).

I Afsnit 1 gennemgår jeg kortfattet begreberne *vægte* og *vejning* samt *klyngeudvælgelse* og *stratificeret udvælgelse*. I Afsnit 2 præsenterer jeg et eksempel, hvor man alene får opgivet en vægt-variabel fra dataudbydernes side og ikke eventuelle andre oplysninger om stikprøven. Dernæst følger i Afsnit 3 et eksempel, hvor der foruden en vægt-variabel også angives klynge- og strata-variabler. I et appendiks behandles en alternativ tilgang til analyse på data med såkaldt *poststratifikation* end den, der benyttes i Afsnit 2 og 3.

1. Vejning af data samt klyngeudvælgelse og stratificering

Vejning på baggrund af stikprøvedesign

Som nævnt i indledningen, er statistisk inferens fra stikprøve til population baseret på *sandsynlighedsudvælgelse*, dvs. hvor man for alle populationens enheder kender sandsynligheden for udvælgelse til stikprøven, men hvor disse sandsynligheder ikke behøver at være ens for alle i populationen, sådan som det er tilfældet ved simpel tilfældig udvælgelse. Det er meget almindeligt, at man i sit design af stikprøven har gode grunde til at planlægge indsamlingsprocessen sådan, at sandsynligheden for udvælgelse er forskellig for forskellige enheder. Valid statistisk inferens under analysen i sådan en situation fordrer imidlertid ofte, at stikprøvens enheder vejes på plads, så fordelingerne kommer til at passe med populationens fordelinger.

En typisk situation i udtræk af stikprøve, hvor der sker ulige udvælgelsessandsynlighed, er, hvis der tilfældigt udvælges adresser, men hvor der kun interviewes én voksen person fra hver adresse. Her vil udvælgelsessandsynligheden være mindre, jo flere voksne der er i husstanden, og for at gøre stikprøven repræsentativ ift. populationen, bør det i analyser sikres, at den enkelte respondent vejes op eller ned i forhold til antal voksne i husstanden.

[1] International Social Survey Programme, European Values Study, World Values Survey og European Social Survey.

Man kan fx også forestille sig en situation, hvor forskere vil undersøge forskelle i holdninger til forskellige issues mellem ledige og folk i arbejde. Hvis man fx blot udtrækker 2.000 respondenter simpelt tilfældigt blandt folk i arbejdsstyrken, vil der komme uhensigtsmæssigt få ledige med til at kunne udtale sig om lediges holdninger, og der ville tilsvarende komme unødvendigt mange folk i arbejde med. Af den grund vil man i sådan en situation vælge en såkaldt *disproportionalt stratificeret udvælgelse*. Populationen opdeles her i to strata, ledige og folk i arbejde, og derefter udtrækkes der fx en simpel tilfældigt udtrukket stikprøve fra hvert strata, men hvor der altså udtrækkes en større stikprøve blandt de ledige. I analyser, hvor der foretages sammenligninger mellem ledige og folk i arbejde, vil det ikke være nødvendigt at veje. Men hvis der i en del af analyserne i projektet skal analyseres på datasættet, som om det var et repræsentativt billede på populationen, bør den enkelte arbejdsløse veje mindre i analysen end den enkelte, der er i arbejde. Hvis man *ikke* vejer sine analyser, og hvis vi siger, at de arbejdsløse har langt mere positive holdninger til et eller andet spørgsmål end dem i arbejde, så vil man få et forkert – alt for positivt – billede af populationens gennemsnitlige holdning til det spørgsmål.

For at kunne foretage vejning i analyserne, skal der først konstrueres en vægt-variabel i datasættet, og hvis der alene skal vejes ift. en designmæssig stikprøveskævhed, kaldes sådan en variabel normalt for en *designvægt*. En klassisk måde at beregne vægten på for den enkelte enhed i stikprøven er som den inverse af inklusionssandsynligheden, forstået som den reciprokke værdi af denne, dvs. én over inklusionssandsynligheden. Hvis vi fx siger, at der i eksemplet ovenfor med oversamplingen af arbejdsløse er 80.000 arbejdsløse i populationen, og vi gerne vil inkludere 1.000 af dem i stikprøven, så har hver arbejdsløs en inklusionssandsynlighed på 1.000/80.000 = 0,0125. Og hvis man benytter sandsynlighedsvægtning efterfølgende, så kan de enkelte enheders vægt bestemmes som 1/0,0125 = 80. Hvis vi tilsvarende vil inkludere 1.000 respondenter blandt folk i arbejde, og der i alt er 4.000.000 i arbejde i populationen, så er inklusionssandsynligheden for hver af disse 1.000/4.000.000 = 0,00025, og deres vægt i analyser med et repræsentativt billede af populationen bliver 1/0,00025 = 4000. De to vægte for henholdsvis de arbejdsløse og dem i arbejde vil alternativt kunne beregnes som N/n, dvs. henholdsvis 80.000/1.000 og 4.000.000/1.000. Vægtene vil med sådanne beregninger summere til populationens tal. I et tidssvarende statistikprogram bliver standardfejl og tilhørende p-værdier dog beregnet ud fra de analyseenheder, der reelt er i stikprøven. Går man blot et årti eller to tilbage i tiden, var det almindeligt lade statistikprogrammerne gennemføre vægtede analyser på en noget mere simpel måde, hvor man statistisk set foregav at have lige så mange respondenter, som vægtene summerede til[2]; og for at undgå alt for store fejl i sikkerhedsestimeringen, beregnede man som oftest vægtene sådan, at de summerede til stikprøvestørrelsen i stedet for til populationens størrelse (evt. re-skalerede statistikprogrammet selv automatisk hertil). Det kan gøres ved at dividere andel i population med andel i stikprøve i stedet for at benytte frekvenserne. De ledige udgør en andel på cirka 0,0196 (80.000/4.080.000) af populationen, mens de udgør halvdelen af stikprøven, så vægten for ledige vil blive lig med 0,0392 (0,0196/0,5). Tilsvarende vil

[2] Det bør tilføjes hertil, at det i nogle situationer kan være helt legitimt at benytte sådan en type vejning, nemlig hvor man har et aggregeret datasæt, hvor fx en analyseenhed med vægten 50 angiver, at der reelt findes 50 af disse enheder – med en engelsk term såkaldt *frequency weighting*.

vægten for folk i arbejde blive lig med 1,9608 (0,9804/0,5) Man ser stadigvæk ofte den metode i dag, men for en valid estimering af statistisk signifikans ifm. analyser er det ligegyldigt. Det vigtige her er blot, at *forholdene* mellem vægtene er korrekte.[3]

Vejning på baggrund af skævt bortfald

I ovenstående er beregning af vægtvariabel og analyse med vejning beskrevet ud fra, at stikprøvens skævhed i forhold til populationen er sket på baggrund af en bevidst strategi i designet af stikprøveudtrækket, såkaldt designvejning. Vægte kan imidlertid også være konstrueret på baggrund af et ikke planlagt *bortfald* fra bruttostikprøve til nettostikprøve, hvor en del af personerne, der er udtrukket til stikprøven, af den ene eller anden årsag ikke besvarer spørgeskemaet. Hvis der har været et anseeligt bortfald, sådan som det i dag kan forventes, og man alene benytter en designvægt, der vejer på plads ud fra brutto-stikprøven, så vil der implicit ligge en forventning om, at bortfaldet er sket helt tilfældigt blandt de udtrukne individer, såkaldt *Missing Completely At Random* (MCAR)[4]. Det er som regel en ret så urealistisk forventning. En lidt bedre løsning vil som regel være at bruge en eller flere hjælpevariabler – på engelsk *auxiliary variables* – til at beregne nogle forventeligt mere valide vægte. Det er variabler, hvor man kender populations-fordelingerne. Eksempelvis kan der have været et større frafald blandt unge, herunder fx især unge mænd. Og hvis man kender populationens fordeling på kombinationer af alder og køn, kan man efter indsamlingen undersøge, hvor godt fordelingen i *stikprøven* på alder og køn svarer til *populationens* fordeling på samme.

I praksis vil man i den situation opdele alder i en række intervaller. Hvis man fx opdeler i seks intervaller og kombinerer dem med mænd og kvinder, får man 12 grupper, som kaldes for *strata*, jævnfør ovenfor Det gør man i både populationen og i stikprøven, og man ser dernæst på, hvor stor en andel de enkelte strata udgør af henholdsvis population og stikprøve. Det er et eksempel på en såkaldt bortfaldsanalyse. Det skal her tilføjes, at hvis der også i designfasen har været et planlagt skævt udtræk til brutto-stikprøven, og at der derfor er genereret en designvægt, skal sammenligningen af fordelinger i population og stikprøve baseres på en designvejet fordeling i stikprøven.

Hvis der er mere end marginale forskelle i fordelingerne mellem population og stikprøve i de valgte strata, kan man vælge at beregne en vægtvariabel, der er baseret på disse forskelle. For hvert stratum kan vægten fx sættes lig med andel i population divideret med andel i stikprøve, sådan at vægtene summerer til stikprøvestørrelsen. En sådan vægt kaldes for en *poststratificeringsvægt*, fordi det ikke er noget, der foregår i designfasen, men derimod efter indsamlingen. Og hvis der findes både en poststratificeringsvægt og en designvægt, skal den endelige vægtvariabel sættes lig med de to vægte multipliceret med hinanden.

---

[3] At den gammeldags metode til at lave vejede analyser kan give anledning til sågar grov fejlestimering af standardfejl og dermed også p-værdier, kan let vises gennem et simpelt eksempel. Hvis der i en population er 50 pct. mænd og 50 pct. kvinder, og vi i en stikprøve kun udtrækker 2 mænd og 98 kvinder, så ville den enkelte mands vægt være lig med 25, mens kvindernes vægt ville være omtrent 0,51. Vi kunne så fx lave nogle vældig fine analyser af holdningsforskelle mellem mænd og kvinder, hvor der sagtens kunne vise sig statistisk signifikante forskelle, men det er også åbenlyst, at det ikke bør kunne lade sig gøre at konkludere vedrørende forhold i populationen, når vi reelt kun har to mænd med i stikprøven.

[4] Se herom samt tilstødende begreber, der kort omtales i dette afsnit i fx Cobben (2009) kap. 1.

Denne type af vejning vil som regel være bedre end at analysere på uvejede eller blot designvejede data. Her er dog stadigvæk en implicit forventning om såkaldt *Missing At Random* (MAR), hvilket betyder, at man forventer et tilfældigt bortfald *inden* for de enkelte strata, bortfaldet er blot større i nogle strata end i andre. I eksemplet ovenfor med et større bortfald blandt især unge mænd, er spørgsmålet altså, om de unge mænd, der ikke har besvaret skemaet, på alle væsentlige parametre ligner de unge, der *har* besvaret det. Hvis spørgeskemaet fx handler om politik, og de unge generelt er mindre interesseret i politik end de ældre, kan *det* jo være en væsentlig årsag til det forholdsvis store bortfald blandt de unge mænd. Sådan en situation kaldes for *Not Missing At Random* (NMAR)[5]. Og her hjælper det ikke rigtig noget at veje de unge mænd, der har besvaret skemaet, op. De unge mænd, der i den situation vejes op, vil ikke være repræsentative for de unge mænd i populationen. De vil være mere politisk interesserede, og derfor vil der stadigvæk efter vejning være et skævt billede af den politiske interesse i populationen.

En analysevægt kan som nævnt være en *kombination* af design- og poststratifikationsvægt. I ESS data får man fx opgjort både en design-vægt og en kombineret design- og post-stratifikationsvægt, hvor der i sidstnævnte altså vægtes for både den designmæssige planlagte skævhed og en ikke planlagt skævhed pga. bortfald. Hvis man alene bruger design-vægten, så vil der implicit ligge en MCAR-forventning, mens der vil ligge en MAR forventning, hvis man benytter en kombineret design- og poststratifikations-vægt.

Den ovenfor gennemgåede metode til vejning efter skævt bortfald er det, man kan kalde for den traditionelle metode. Der findes imidlertid andre metoder, primært den der med en engelsk term kaldes for *propensity score* vejning. På et datasæt bestående af samtlige personer i brutto-stikprøven, estimeres der her en sandsynlighed/propensity for besvarelse af spørgeskemaet, altså deltagelse i netto-stikprøven. Hertil benyttes som oftest en logistisk regression, hvor den afhængige variabel er lig med 1, hvis personen er med i nettostikprøven, og 0 i modsat fald. For personer, hvor der estimeres en lille sandsynlighed for deltagelse i netto-stikprøve, beregnes der en stor vægt-værdi, mens der omvendt for personer med stor sandsynlighed for inklusion i netto-stikprøven beregnes en lille vægt-værdi. Metoden er mere fleksibel ift. at kunne medtage intervalskallerede auxiliary variabler, ligesom der kan medtages flere auxiliary variabler i estimeringen end ved traditionel strata-opdeling. Hvor god den endelige vægt er, afhænger bl.a. af, hvor godt modellen specificeres. Propensity metoden beskrives ikke yderligere i denne artikel.

Klyngeudvælgelse og stratificering

Ved download af sekundære survey-data vil der fra dataudbydernes side nogle gange ligge information om *klyngeudvælgelse* og *stratifikation*, hvor der så – ud over vægte – kan være inkluderet variabler til identifikation af klynger og strata.

At man foretager klyngeudvælgelse betyder, at man først inddeler populationen i en serie klynger, fx geografiske områder som kommuner, og dernæst udvælger et vist antal kommuner til sin stikprøve. Hvis det drejer sig om en survey-undersøgelse, vil man i andet trin udvælge en række personer fra hver af de valgte kommuner. Både klynge- og individudvælgelsen foregår ved en eller anden form for sandsynlighedsudvælgelse, dvs.

[5] Betegnelsen veksler i litteraturen mellem *Not Missing At Random* (NMAR) og *Missing Not At Random* (MNAR).

hvor man kender sandsynligheden for det enkelte individs udvælgelse. Som oftest ligger der økonomiske/ressourcemæssige hensyn bag klyngeudvælgelse, ikke mindst hvis der planlægges personlige interviews med besøg i respondenternes eget hjem.

Klyngeudvælgelse går imidlertid ud over den statistiske sikkerhed. Det er let at forstå, hvis man forestiller sig en ekstrem og ikke realistisk situation, hvor man tilfældigt udvælger én kommune blandt danske kommuner, hvorefter man udvælger 1.000 personer fra den valgte kommune. Næsten uanset hvilken kommune, man tilfældigt udvælger, vil det give nogle meget skæve resultater[6]. Hvis man i stedet udvælger to kommuner, og fra hver af disse udvælger 500 personer til sin stikprøve, vil der være mindre sandsynlighed for et skævt resultat i analyserne, men det vil stadigvæk være ret så usikkert ift. at kunne konkludere om forhold i hele populationen samlet set. Jo flere kommuner, der vælges (med samme stikprøvestørrelse alt i alt), des mindre standardfejl og sikkerhedsmargin på estimaterne. Det er altså ofte en afvejning mellem økonomi og statistisk sikkerhed.[7]

I forhold til simpel tilfældig udvælgelse kan man imidlertid ved klyngeudvælgelse genvinde noget af sikkerheden i estimaterne, hvis man samtidigt foretager *stratificeret* udvælgelse. Ved stratificeret udvælgelse opdeler man populationen i en serie strata under designfasen. Ovenfor nævnte jeg et eksempel, hvor der i en population af personer i arbejdsstyrken blev opdelt på to strata, ét for ledige og et andet for folk i arbejde, og fra begge disse strata blev der valgt individer til stikprøven. Til forskel fra klyngeudvælgelse, hvor der kun udvælges en del af klyngerne, hvorfra der så i andet trin udvælges enheder til stikprøven, udvælges der ved stratificering enheder fra samtlige strata. Som oftest vil der være flere end to strata. Til eksempel kan disse bestå af kombinationer af køn, alder og uddannelse, hvor man så inden for hvert stratum udtrækker en tilfældig stikprøve. Alder er jo intervalskaleret, og for at kunne opdele i strata kræves det, at man alene benytter faktorer, der består af kategorier. Derfor inddeles alder forinden i en serie aldersintervaller.

En kombination af klyngeudvælgelse og stratificeret udvælgelse kan fx foregå ved, at man simpelt tilfældigt udvælger en stikprøve af kommuner og derefter foretager stratificeret udvælgelse inden for de udvalgte kommuner. Men man kan også koble en stratificeret udvælgelse på selve klyngeudvælgelsen, fx sådan at man trækker en stikprøve blandt små kommuner for sig og af store kommuner for sig. Og man kan lave mange forskellige typer af kombinationer og trin. Fx kunne man også i en klyngeudvælgelse sige, at man partout vil have kommunerne med de fire største byer med i stikprøven, hvorefter resten af udvælgelsen foregår ved sandsynlighedsudvælgelse på den ene eller anden vis.

---

[6] Som et kuriosum kan det dog nævnes, at der i særlige tilfælde kan findes undtagelser herpå, fx hvor en valgkreds ved gentagne folketingsvalg viser tendenser til at ligne landsresultatet. Hvis der fra denne valgkreds er tidlige resultater, kan disse (i en snæver vending) bruges som indikation på landsresultatet. Inden for fiktionsverdenen kendes et lignende eksempel fra filmen *Magic Town* med James Stewart i hovedrollen.

[7] Der kan dog også være særlige årsager til at foretage klyngeudtræk, som ikke har noget med økonomi at gøre. Hvis man fx har hypoteser angående effekter på forskellige niveauer, som fx individ- og kommune-niveau, kan det være en fordel at klyngeudtrække kommuner med et antal individer i hver, sådan at datasættet egner sig til såkaldt multilevel analyse.

Stratificeret udvælgelse giver typisk en større grad af sikkerhed i estimaterne end simpel tilfældig udvælgelse. For det første vil der i bruttostikprøven ikke forekomme stikprøve-skævhed på de parametre, som man benytter i konstruktionen af strataene, hvilket der i simpel tilfældig udvælgelse altid vil forekomme i et vist omfang pr. tilfældighed. For det andet vil man kunne vinde noget sikkerhed i estimeringen, hvis der inden for de enkelte strata er en mindre variation på de variabler, man har i fokus, end der er på pågældende variabler alt i alt i stikprøven. Og jo mere strataene adskiller sig fra hinanden på de variabler, man er interesseret i at måle på, fx politiske holdninger, des mere vil man vinde ift. simpel tilfældig stikprøveudvælgelse. Fx vil standardfejlen til et gennemsnit på en variabel være beregnet ved en kombination af variansen *inden* for de enkelte strata. Hvor god en stratificeret udvælgelse, man er i stand til at designe, afhænger altså af, dels hvilken information man har om variabler knyttet til de enkelte individer i populationen på forhånd, dels hvor velunderbyggede hypoteser man har om korrelation mellem disse og de afhængige variabler i analyserne.

Da både klyngeudvælgelse og stratificeret udvælgelse påvirker estimaternes sikkerhed, bør oplysninger herom optimalt set indregnes i analyserne, såfremt de er opgivet fra dataudbydernes side, selvom det væsentligste i de fleste tilfælde nok er selve vejningen. Disse muligheder findes i dag i alle de gængse statistikprogrammer som SAS, SPSS, Stata og R. I Afsnit 3 vil jeg gennemgå et konkret eksempel i R, hvor der er oplysninger om både vægte, klynger og strata. I Afsnit 2 herunder vil jeg imidlertid starte med et eksempel, hvor der alene er opgivet en vægtvariabel, hvilket er en ret typisk situation for forskere og studerende, der gennemfører analyser på sekundære surveydata. I et appendiks vil jeg diskutere eksemplet fra Afsnit 2 yderligere og prøve at tilgå det i R med en lidt mere udspecificeret tilgang, men – viser det sig i eksemplet – i bund og grund med samme resultater i praksis.

2. Eksempel 1: Vejning af ISSP-data i R med pakken `survey`

Jeg har fra AAU Surveybanken[8] downloadet de danske data fra ISSP-undersøgelsen fra 2017, og i R har jeg kaldt datasættet for "ISSP17". Stikprøven var simpelt tilfældigt udtrukket fra CPR, så der er ikke behov for designvejning. Som det dog ofte hænder, viste det sig for Rambøll, der stod for dataindsamlingen, at der var et betydeligt bortfald af respondenter, og at bortfaldet ikke var sket tilfældigt. Man fandt i bortfaldsanalyser, at stikprøven var skæv på alder- og kønsstrata, og der blev konstrueret en post-stratifikationsvægt, der i analyser kunne veje respondenter op i de strata, hvor der var en for lille andel ift. fordelingen i populationen, og omvendt veje respondenter ned i de strata, hvor der var en for stor andel i forhold til fordelingen i populationen. Generelt var der en underrepræsentation af yngre generationer og mænd. Vægtvariablen, som i datasættet hedder `v153` summerer til stikprøvestørrelsen, dvs. at den har en gennemsnitlig værdi på 1. Som nævnt i Afsnit 1 er det for den statistiske inferens ikke vigtigt, at vægten er skaleret sådan. Den kunne fx også have været skaleret til at summere til populationsstørrelsen. Det vigtige er *forholdet* mellem respondenternes vægt-værdi.

[8] https://www.surveybanken.aau.dk/

Jeg står med disse ISSP-data i en lidt uheldig (men efterhånden helt almindelig) situation, hvor der er et betydeligt og skævt bortfald. Hvis der er tale om en MAR-situation, vil en vejning på køn- og alders-strata kunne rette en del op på skævhederne i stikprøven, sådan at analyseresultaterne får større ekstern validitet. Men jeg har ingen garanti for, at det forholder sig sådan, og jeg kan ikke undersøge det. Jeg vælger her at tro på, at en vejning i det mindste vil forøge validiteten en del, altså at bortfaldet inden for de enkelte strata ikke afviger meget fra en MAR-situation.

Først installerer jeg `survey`-pakken, og dernæst specificerer jeg vægtvariablen med kommandoen `svydesign`, og jeg gemmer specifikationen under navnet `design1`:

```
install.packages("survey")
library("survey")
design1 <- svydesign(id = ~1, weights = ~V153, data = ISSP17)
```

Jeg har nu en simpel designbeskrivelse, `design1`, som *linker* til datasættet `ISSP17`. Inde i parentesen i kommandoen findes der tre oplysninger. Den første del er påkrævet, og den drejer sig om, hvorvidt der er foretaget klyngeudvælgelse. Ifald der var det, skulle der her angives en variabel til identifikation af klyngerne. I dette datasæt er der ikke klyngeudvælgelse, og i den situation skal der blot stå "`id = ~1`" (eller "0" i stedet for "1"[9]). I den sidste af de tre dele i parentesen angives blot det datasæt, som beskrivelsen skal knyttes til. Hvis der alene var disse to oplysninger, ville efterfølgende analyser med survey-pakken give samme resultater som uden brug heraf. Nu har jeg imidlertid også angivet en vægtvariabel med teksten "`weights = ~V153`".

Herefter kan jeg benytte forskellige `svy`-analysekommandoer, som ligner de kommandoer, man er vant til at bruge, fx kommandoerne `table`, `mean` og `glm`. Herunder viser jeg frekvenser samt beregning af gennemsnit for variablen `V39` i datasættet, dels uvægtet, dels vægtet. Variablens værdier er svar på et spørgsmål om tillid til domstolene med svarkategorier fra 0 til 10. Skalaen betragtes her som intervalskaleret.

```
table(ISSP17$V39)

  0   1   2   3   4   5   6   7   8   9  10
 14   8  14  26  25  84  46  81 242 224 203

mean(ISSP17$V39, na.rm = TRUE)

[1] 7.713547

svytable(~V39, round = TRUE, Design1)

V39
  0   1   2   3   4   5   6   7   8   9  10
 14   9  15  27  25  86  46  84 239 224 198

svymean(~V39, na.rm = TRUE, Design1)

      mean     SE
V39 7.6794 0.0739
```

[9] Og jeg kunne også angive en unik id-variabel, som vil give samme resultat (se herom i Afsnit 3).

I `svy`-analysekommandoerne ovenfor er det vigtigt at angive design-navnet i stedet for navnet på data. Det er ikke nødvendigt at angive data, da designet linker hertil. Læg også mærke til, at der skal en tilde foran variabelnavnet i analyserne.

Det vægtede gennemsnit på tillid til domstolene er en anelse lavere end det uvægtede. Det er udtryk for, at der i stikprøven er kommet for mange ældre med ift. de unge, og at de ældre generelt har større tillid end de unge. Når man så vejer de ældre lidt ned og de unge lidt op, bliver gennemsnittet på variablen lidt lavere, og det er (forhåbentligt) et lidt mere validt estimat af populationens gennemsnit på variablen, selvom det må siges at være i småtingsafdelingen her.

Jeg kunne nu vælge at gennemgå en lang række `svy`-kommandoer, der kunne være relevante i eksemplet. Jeg vil imidlertid nøjes med endnu ét eksempel, nemlig på en vejet lineær regression. Øvrige analysekommandoer, der måtte være behov for, vil jeg lade være op til selvstudie. På hjemmesiden "Survey data analysis with R"[10] fra UCLA vises en lang række eksempler på `svy`-analysekommandoer.

Regression i R på vejede data

Jeg går nu videre fra den indledende analyse ovenfor til en multipel lineær regressionsanalyse, dels uvejet, dels vejet. Herefter viser jeg, hvordan man med `marginaleffects`-pakken kan præsentere udvalgte resultater fra den vejede analyse grafisk.

Først vises en uvejet lineær regression, hvor variablerne `V85K` (dummy for kvinde), `V86` (alder i hele år) og `studeks` (dummy for ungdomsuddannelse) prædikterer `V39` (tillid til domstolene):

```
lm1_uw <- lm(V39 ~ V85K + V86 + studeks,
             na.action = na.omit,
             data = ISSP17)
summary(lm1_uw)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.850909   0.275505  21.237  < 2e-16 ***
V85K        -0.274304   0.141320  -1.941   0.0526 .
V86          0.030368   0.004536   6.694 3.76e-11 ***
studeks      1.071541   0.152138   7.043 3.67e-12 ***
```

Der ses en negativ effekt fra kvinde-dummien, dog kun knapt nok statistisk signifikant på 0,05 niveau samt højsignifikante positive effekter fra alder og ungdomsuddannelse. Jeg undersøger nu, hvordan det ser ud på de vejede data:

```
lm1_w <- svyglm(V39 ~ V85K + V86 + studeks,
                na.action = na.omit,
                design = Design1)
summary(lm1_w)
```

[10] https://stats.oarc.ucla.edu/r/seminars/survey-data-analysis-with-r/

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.790661   0.321484  18.012  < 2e-16 ***
V85K        -0.324883   0.143728  -2.260    0.024 *
V86          0.031326   0.005035   6.222 7.44e-10 ***
studeks      1.142792   0.165208   6.917 8.60e-12 ***
```

Kønsvariablen er nu lidt stærkere, og den er signifikant på 0,05 niveau. Men alt i alt ligger resultaterne fra henholds uvejet og vejet analyse meget tæt på hinanden, ligesom det var tilfældet i den univariate analyse ovenfor.

Læg i øvrigt mærke til, at jeg i den vægtede regressionsanalyse benytter en `svy`-udgave af `glm`-kommandoen i R, da der ikke findes en `svy`-udgave af `lm`-kommandoen. Imidlertid er default med `glm` en lineær regression. Der kommer ikke R^2 med i udskriften ved en `summary` af `glm` eller `svyglm`, men den kan fås fx via `jtools`-pakken (Long, 2022). Hvis denne installeres, kan der efter `svyglm`-kommandoen køres en `summ`-kommando (hvor jeg desuden beder om tre cifre efter decimaladskiller):

```
summ(lm1_w, digits = 3)
```

```
MODEL INFO:
Observations: 926
Dependent Variable: V39
Type: Survey-weighted linear regression

MODEL FIT:
R² = 0.078
Adj. R² = 0.075

Standard errors: Robust
------------------------------------------------
                      Est.    S.E.   t val.      p
----------------- -------- ------- -------- -------
(Intercept)           5.791   0.321   18.012   0.000
V85K                 -0.325   0.144   -2.260   0.024
V86                   0.031   0.005    6.222   0.000
studeks               1.143   0.165    6.917   0.000
------------------------------------------------
```

I forbindelse med ovenstående regressionsanalyse skal der lige gøres en tilføjelse. Én af årsagerne, til at resultater fra uvejede regressionsanalysens ofte ligger tæt op ad de vejede, er, at de faktorer, der benyttes til strata-opdeling og vægtberegning, er medtaget i regressionsanlyserne som uafhængige variabler. Men for det første vil der også ofte estimeres regressionsmodeller, hvor disse faktorer ikke er inkluderet som uafhængige variabler. For det andet er det ikke givet, at de inkluderes i samme form som i strataopdelingen. I regressionsanalyserne ovenfor er begge faktorer, køn og alder, ganske vist inkluderet som uafhængige variabler, men ikke i samme form som strataene. I så fald skulle alder have været som faktorvariabel med seks kategorier, og der skulle have været specificeret interaktion mellem variablerne for køn og alder. Uanset hvad, er der dog her ikke den store forskel mellem de vejede og de uvejede resultater.

Grafisk visning af effekter fra de vejede analyseresultater

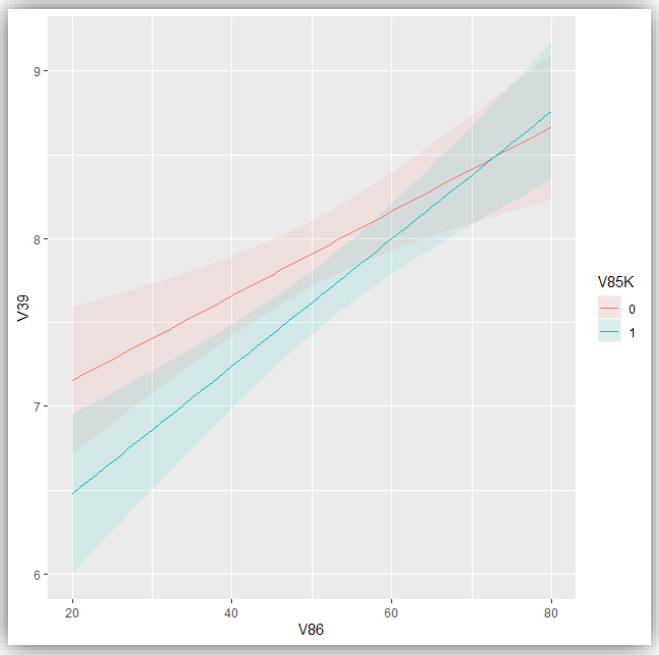
Ofte vil man have behov for at vise udvalgte, kontrollerede effekter grafisk pba. de vejede regressionsanalyser. Hertil kan man fx benytte pakken `marginaleffects`, som også understøtter kommandoen `svyglm`. Grafisk visning af effeker i forbindelse med lineær regression er især relevant, når der inkluderes interaktionseffekter og/eller ikke-lineære effekter som fx logittransformrede uafhængige variabler eller inklusion af en uafhængig variabel i både 1. og 2. potens (dvs. den uafhængige variabel + et kvadreret led). Jeg genererer herunder først en vejet lineær model med interaktion mellem køn og alder, samt derefter en vejet lineær model med tillæg af kvadreret alder, begge analyser med grafisk visning af effekter efterfølgende. De uafhængige variabler, der ikke nævnes eksplicit i `condition`-option, sættes til deres gennemsnit (modus ved faktor-variabel). Modellerne vises blot for eksemplet skyld, for ingen af dem er statistisk set bedre end den mere simple model vist ovenfor[11].

```
#Interaktion mellem køn og alder:
lm2_w <- svyglm(V39 ~ V85K * V86 + studeks,
                na.action = na.omit,
                design = Design1)
summary(lm2_w)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.076111   0.381650  15.921  < 2e-16 ***
V85K        -0.929936   0.471049  -1.974 0.048660 *
V86          0.025215   0.006583   3.830 0.000137 ***
studeks      1.154333   0.165250   6.985 5.44e-12 ***
V85K:V86     0.012788   0.008834   1.448 0.148078
```

```
plot_predictions(
  lm2_w,
  condition = c("V86", "V85K"))
```

[11] Jeg har afprøvet forskellige andre modifikationer med disse tre uafhængige variabler i grundstrukturen, men ingen viste sig bedre end den viste med de tre hovedeffekter.



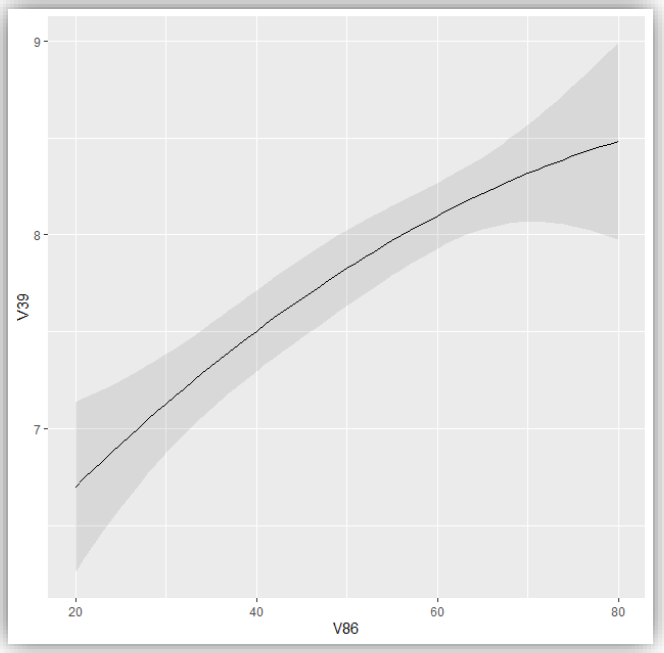
Afslutningsvis viser jeg herunder eksempel på grafisk visning af en kvadratisk effekt fra alder. Læg mærke til, at jeg *ikke* først beregner en kvadreret version af variablen `V86`, men derimod får den beregnet inde i regressionsformlen. Det er vigtigt, hvis alderseffekten efterfølgende skal vises grafisk på en nem måde.[12]

```
lm3_w <- svyglm(V39 ~ V85K + V86 + I(V86^2) + studeks, na.action =
na.omit, design = Design1)
summary(lm4_w)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.2937511  0.6319290   8.377  < 2e-16 ***
V85K        -0.3227231  0.1436103  -2.247   0.0249 *
V86          0.0558641  0.0266412   2.097   0.0363 *
I(V86^2)    -0.0002616  0.0002726  -0.960   0.3375
studeks      1.1303993  0.1658905   6.814 1.71e-11 ***

plot_predictions(
  lm3_w,
  condition = "V86")
```

[12] Skriv "help(I)" i R-konsollen for yderligere forklaring.



Ved lineære modeller som her er det som regel ret simpelt at specificere grafer med `plot_predictions`.[13] Ved logit- og probit-modeller kan det lidt mere kompliceret. Her anbefales det at plotte såkaldt *counterfactual predictions*, der specificeres med et `newdata`-argument i `plot_predictions`-kommandoen (se evt. Lolle, 2024 herom).

Analyse på udsnit af stikprøvens enheder

Man kan udmærket med `svy`-kommandoer analysere på subsets af data, som man plejer at gøre, men der vil være mere "power" i at foretage svy-subsetting i stedet for. Ved svy-subsettting fås samme estimater af parameterne som ved normal subsetting, men som regel vil der være en bedre sikkerhed i form af mindre standardfejl, fordi også dataenheder, der ikke er med i subsettet, bruges i beregningen af standardfejl. Herunder viser jeg en normal subset-procedure efterfulgt af en `svymean`-kommando samt derefter samme analyse med en `svy`-subsetting. I den normale subsetting laver jeg først en tilskæring af datasættet, her blot respondenter i hovedstaden. Variablen `V157` angiver region, og hovedstadsområdet har koden 1084. Herefter skriver jeg en ny `svydesign`-kommando for dette datasæt, og endelig bruger jeg så `svymean`-kommandoen til at få printet gennemsnit og standardfejl ud på variablen `V39`.

```
ISSP17KBH <- ISSP17 %>%
  filter(V157 == 1084)
DesignKBH <- svydesign(
  id = ~1,
  weights = ~V153,
  data = ISSP17KBH)
svymean(~V39, na = TRUE, DesignKBH)

      mean     SE
V39 8.0849 0.1146
```

[13] Se evt. https://marginaleffects.com/ for flere muligheder.

I stedet for at skære datasættet til som ovenfor, benytter jeg nu kommandoen `subset` til at subsette `Design1`.

```
Design_svyKBH <- subset(Design1, V157 == 1084)
svymean(~V39, na = TRUE, Design_svyKBH)

      mean     SE
V39 8.0849 0.1145
```

Hovedgevinsten her er nok, at det er kortere at kode med svy-tilgangen, for gevinsten i form af mindre standardfejl er ekstremt lille. Med normal tilgang er standardfejlen 0,1146, mens den er 0,1145 med `svy`-tilgangen.

3. Eksempel 2: ESS data med vejning samt specifikation af klynger og strata i R med pakken `survey`

Designbeskrivelsen bliver lidt mere kompliceret i takt med, at selve stikprøveudtrækket er designet på en mere kompleks facon. I fx European Social Survey foreligger der oplysninger om dels forskellige vægte, dels strata- og klyngeudvælgelse, og variablerne herfor findes for samtlige lande i datasættet. I nogle lande er der først foretaget klynge-udvælgelse, fx af kommuner, hvor inklusionssandsynligheden for en kommune afhænger af antallet af indbyggere i kommunen plus evt. andre aspekter for at gøre den endelige stikprøve så repræsentativ som muligt. Det, at inklusionssandsynligheden er større for kommuner med stort befolkningstal, kaldes for en proportionalt stratificeret klyngeudvælgelse[14]. Overordnet set er der dernæst foretaget simpelt tilfældigt eller systematisk udvælgelse af respondenter. I nogle lande er det personer, der udvælges direkte, mens det i andre er husstande, hvor der så vælges én person ud fra hver husstand. Hvis det sidste er tilfældet, får det betydning for vægtenes konstruktion, så der ikke bliver skævhed i analyserne på husstandsstørrelse. I andre lande er der blot ét enkelt trin i udvælgelsen, dvs. ingen klyngeudvælgelse til at starte med. Det gælder fx Danmark, hvor CPR er benyttet til simpel tilfældig udvælgelse af personer. Der er dog stadigvæk en slags klyngevariabel, eller rettere det, der kaldes for en PSU (Primary Sampling Unit). I Danmark er dette blot respondentens unikke id-nummer, da individerne er PSU i Danmark. Mht. strata regnes Danmark endvidere som ét stratum, altså blot med en konstant på strata-variablen.

Hele det samlede ESS datasæt er overordnet designet sådan, at den samme formular for surveydesign kan bruges, uanset hvor mange af landene man benytter i sine analyser.[15] Der er imidlertid flere vægt-variabler i datasættet, og valg af specifik vægtvariabel afhænger af, om man har ét samlet geografisk område bestående af flere lande som den population, man vil inferere til, eller om man vil lave enten enkeltlandsundersøgelse eller lande*komparativ* undersøgelse. Er det en samling af lande, der er populationen, skal der

[14] I nogle lande var det med "tilbagelæggelse", så kommuner kunne udvælges mere end én gang.
[15] I hvert fald uden at begå nævneværdige fejl i analysernes sikkerhedsestimationer. Umiddelbart vurderer jeg, at der i nogle lande vil være nogle spidsfindigheder, der kunne komme med i en spcificering af designet, men som ikke kommer med, fordi hele datasættet skal kunne beskrives på samme måde.

vægtes med en variabel, der justerer for antallet af indbyggere i de enkelte lande, mens dette ikke er tilfældet ved fx landekomparativ analyse.

Jeg gennemgår herunder kortfattet et eksempel på opsætning og analyse på et landeudsnit fra ESS round 9, og hvor tanken er at foretage landekomparativ analyse efterfølgende. Jeg medtager data fra Danmark, Tyskland og Storbritannien, og datafilen hedder `ESS9_3lande`.

Først specificerer jeg designet[16]. Variablerne i datasættet, der angiver *Primary Sampling Unit* og *strata*, hedder henholdsvis `psu` og `stratum`. Der findes hele fire forskellige vægtvariabler i datasættet, en designvægt (`dweight`), en kombineret design- og poststratifikationsvægt (`pspwght`), en populationsvægt (`pweight`) samt en kombineret design-, poststratifikations- og populationsvægt (`anweight`). Som oftest vil der skulle benyttes enten `pspwght` eller `anweight`, afhængigt af sigtet med analyserne jævnfør ovenfor. Jeg benytter her `pspwght`, da sigtet er en landekomparativ analyse.

```
design_3lande <- svydesign(ids = ~psu,
                           strata = ~stratum,
                           weights = ~pspwght,
                           data = ESS9_3lande)
```

Herefter foretager jeg en lineær regression med livstilfredshed (`stflife`) som afhængig variabel og tre uafhængige variabler for land (`cntry.f`), køn (`kvinde`) og alder (`agea`). Endvidere specificerer jeg interaktion mellem land og køn, og alderseffekten specificeres med tillæg af alder i anden potens:

```
lm_ess9 <- svyglm(stflife ~ cntry.f * kvinde + agea + I(agea^2),
                  na.action = na.omit,
                  design = design_3lande)
summary(lm_ess9)
```

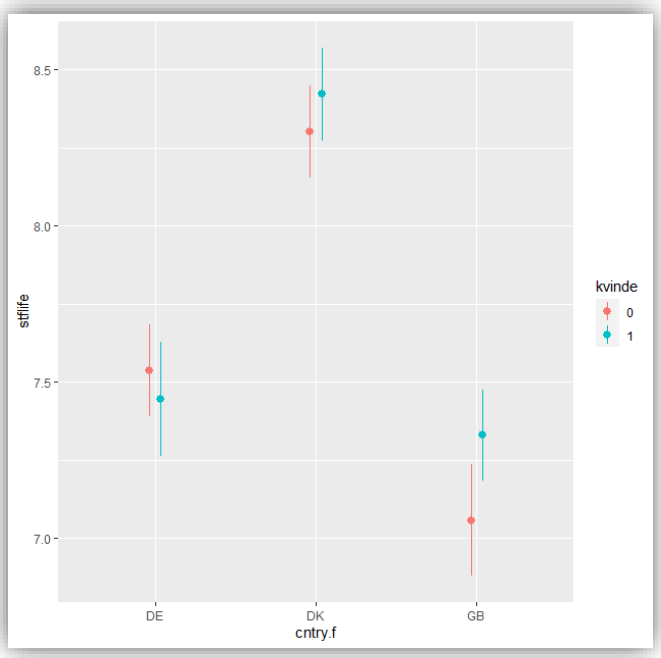
```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.091e+00  1.763e-01  45.900  < 2e-16 ***
cntry.fDK          7.637e-01  9.754e-02   7.830 8.15e-15 ***
cntry.fGB         -4.798e-01  1.081e-01  -4.437 9.66e-06 ***
kvinde            -9.223e-02  1.115e-01  -0.827 0.408236
agea              -2.519e-02  7.861e-03  -3.205 0.001374 **
I(agea^2)          2.816e-04  8.004e-05   3.518 0.000446 ***
cntry.fDK:kvinde   2.114e-01  1.507e-01   1.403 0.160866
cntry.fGB:kvinde   3.643e-01  1.520e-01   2.396 0.016674 *
```

Jeg kunne nu godt begynde at fortolke disse resultater, som fx at danske mænd typisk ligger 0,77 over tyske mænd i livstilfredshed. Det ville ikke være meget vanskeligt. Alligevel vil det være meget nemmere at overskue ved en grafisk præsentation, så jeg laver i stedet for et par grafiske eksempler på præsentation af effekter, nemlig et plot over prædikterede værdier på afhængig variabel for kombinationer af land og køn, samt derefter et plot over alderseffekten:

```
plot_predictions(
```

[16] Se også herom i Kaminska (2023): *Guide to Using Weights and Sample Design* https://stessrelpubprodwe.blob.core.windows.net/data/methodology/ESS_weighting_data_1_2.pdf

```
  lm_ess9,
  condition = c("cntry.f", "kvinde"))
```



Der er tydeligvis forskel mellem landene i livstilfredshed. Men er der forskel mellem de to køn inden for det enkelte land? Her kunne der se ud til at være signifikant forskel mellem kvinder og mænd i Storbritanien, men jeg er ikke helt sikker[17]. Jeg får først R til at prædiktere værdi på afhængig variabel for de to køn for Storbritanien med øvrige variabler i modellen holdt på deres gennemsnit/modus. Jeg bruger kommandoen `predictions` fra `marginaleffects`-pakken. Derefter tester jeg forskellen mellem de to estimater med `hypotheses`-kommandoen, også fra `marginaleffects`.

```
pred <- predictions(
  lm_ess9,
  newdata = datagrid(cntry.f = "GB",
                     kvinde = c(0, 1)))
pred
```

```
Estimate Std. Error  z Pr(>|z|)   S 2.5 % 97.5 % agea cntry.f kvinde
    7.06     0.0905 78   <0.001 Inf  6.88   7.23 50.7      GB      0
    7.33     0.0740 99   <0.001 Inf  7.18   7.47 50.7      GB      1
```

```
hypotheses(pred, hypothesis = "b2 = b1")
```

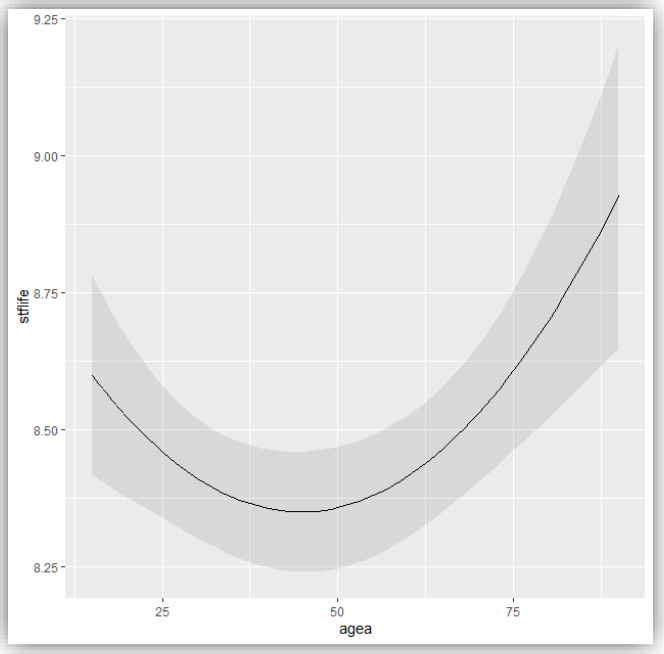
```
  Term Estimate Std. Error    z Pr(>|z|)   S 2.5 % 97.5 %
 b2=b1    0.272      0.103 2.64  0.00833 6.9  0.07  0.474
```

Forskellen mellem de to køn i Storbritannien er 0,27, og den er statistisk signifikant på 0,01 niveau.

[17] En sammenligning mellem de to 95 pct. sikkerhedsintervaller fra plot'et er ikke nogen test med optimal "power".

Afslutningsvis viser jeg herunder også alderseffekten grafisk. Jeg vælger blot at vise den for Danmark (effekten er specificeret til at være ens i de tre lande, så det er alene niveauet mellem landene, der adskiller sig her). Effekten fra alder viser en U-formet kurve, som meget ofte findes ved analyser af livstilfredshed[18].

```
plot_predictions(
  lm_ess9,
  by = "agea",
  newdata = datagrid(
    "agea" = c(15:90),
    "cntry.f" = "DK"))
```



Afslutning

Som det er fremgået af ovenstående, *kan* det at veje analyser samt på anden vis at medregne forskellige aspekter af stikprøvedesignet være forholdsvis enkelt. Med `svydesign`-kommandoen kan designet specificeres, og derefter ligner analyserne dem, vi er vant til at bruge. Det skal dog siges hertil, at der i artiklen er vist eksempler på nogle ikke alt for komplekse specifikationer af stikprøveudtræk. Fx kan man sagtens komme ud for situationer, hvor der skal specificeres klynger/enheder på flere niveauer, evt. med stratificeret udvælgelse på disse, sådan at der fx er to id-variabler og to strata-variabler plus evt. også flere vægt-variabler. Det kan også lade sig gøre at specificere med `svydesign`-kommandoen, men man skal holde tungen lige i munden, og hvis man laver en forkert specifikation, vil samtlige efterfølgende analyseresultater også blive mere eller mindre fejlbehæftede. Ofte vil der dog også i den slags situationer findes vejledning fra dataudbydernes side, om hvordan man kan specificere stikprøven i gængse statistik-

[18] Jeg kunne også bruge `condition` som option ligesom i tidligere eksempler. `newdata` option giver imidlertid samme resultat, og umiddelbart lidt pænere i denne situation.

programmer. Det skal også nævnes, at der er flere mulige "arguments" i kommandoen svydesign, jeg ikke er kommet ind på. For at nævne et enkelt eksempel her, findes der en mulighed for at foretage såkaldt "finite population correction". Man kan tilføje et populationstal, og hvis stikprøven udgør en anseelig andel at populationen, forøges sikkerheden i estimaterne. Som oftest vil stikprøven kun udgøre en ganske lille del at populationen, så den gevinst er yderst marginal. Desuden går man som analytiker som oftest ikke i at kunne udtale sig meget sikkert om en helt konkret population på et givent tidspunkt, men mere på at kunne foretage konklusioner vedrørende nogle systematikker og sociale mekanismer, der bør kunne forklares teoretisk. Det er fx også årsagen til, at de fleste kommunalforskere stadigvæk benytter helt ordinære analysemetoder med signifikansberegninger, selvom de jo meget ofte sidder med datasæt med alle landets kommuner, dvs. hele populationen, fx fra Indenrigs- og Sundhedsministeriets Kommunale Nøgletal[19].

Ud over eventuelle problemer med at få lavet den korrekte specifikation af sit datasæt med `svydesign`-kommandoen er der så også spørgsmålet om, hvorvidt man overhovedet bør benytte vejning, og i givet fald i hvilke typer af analyser der bør vejes.

Som jeg kort var inde på i Afsnit 1, kan poststratificering give en falsk tryghed, fordi man mere eller mindre bevidst som forsker eller studerende antager, at det er en MAR-situation, hvor der inden for de enkelte strata har været tilfældigt bortfald. Så post-stratifikation er i hvert fald ikke er noget columbusæg.

I Afsnit 2 nævnte jeg, at det i multipel regressionsanalyse i nogle situationer ikke er nødvendigt at veje, nemlig hvis de faktorer, hvor der er skævhed ift. fordelingen i populationen (grundet design eller bortfald), er medtaget som uafhængige faktorer i regressionsanalysen. Hvis man sikrer sig det, kan uvejede analyser vise sig mere efficiente end vejede analyser, dvs. have mere power til at spotte effekter. Som Bollen et al (2016) skriver: "If researchers do not weight when appropriate, they risk having biased estimates. Alternatively, when they unneccessarily apply weights, they can create an inefficient estimator without reducing bias." Det er årsagen til, at nogle argumenterer for, at der nok bør vejes i de mere indledende univariate analyser og måske krydstabeller, mens der i fx regressionsanalyser ikke bør vejes. Atter andre argumenterer for, at man bør foretage regressionsanalyse i både vejet og uvejet udgave og derpå efter en sammenligning af resultaterne diskutere, hvad der i situationen er bedst, og som udgangspunkt vil man her som regel håbe på, at der kun er marginale forskelle mellem de to sæt af resultater. I så fald kan man vælge de uvejede analyser samt tilføje, at der ikke er fundet bemærkelsesværdige forskelle mellem de vejede og uvejede resultater. I det hele taget er spørgsmålet om vejning et af de mest diskuterede, ikke blot blandt samfundsforskere, men også blandt økonometrikere og statistikere.

Et sidste aspekt, der bør nævnes, er, at der ganske vist findes rigtig mange `svy`-analysekommandoer, men at man *kan* komme ud for analysemetoder, hvor der ikke findes en tilsvarende `svy`-version, og man kan også komme ud for, at der findes en `svy`-version, men at fx en pakke som `marginaleffects` ikke understøtter denne `svy`-kommando. Man kan fx udmærket foretage både binær logistisk regression, ordinal

[19] https://www.noegletal.dk/noegletal/ntStart.html

logistisk regression og multinomial logistisk regression med `svy`-kommandoer, men af disse tre analysemetoder er det alene den binære logistiske regression, der i skrivende stund understøttes af `marginaleffects`-pakken, hvad angår svy-versionerne.

Appendiks: Poststratificering genbesøgt

I Afsnit 2 gennemgik jeg et eksempel med den danske del af ISSP datasættet fra 2017. I en teknisk rapport til datasættet[20] blev det oplyst, at respondenterne blev udtrukket simpelt tilfældigt fra CPR, samt at man i nettostikprøven, dvs. datasættet, havde fundet skævhed på kombinationer af køn og alder ift. populationen. Rambøll havde pba. skævheden genereret en vægt baseret på køn og aldersfordelingen, sådan at brug af vægten i analyser ville gøre datasættet mere repræsentativt, underforstået med forudsætning om en ikke for stor afvigelse fra en missing at random (MAR).

I Afsnit 2 benyttede jeg blot vægtvariablen i kommandoen `svydesign`. Imidlertid blev der i den tekniske rapport også oplyst hvilke strata, der var lavet vægtvariabel ud fra. Mænd og kvinder var blevet opdelt på hver seks angivne aldersintervaller, sådan at der var i alt 12 strata. Der var ganske vist ikke angivet nogen strata-variabel, men jeg konstruerede selv en strata-variabel ud fra de 12 forskellige værdier i vægt-variablen, og dernæst kunne jeg fra Danmarks Statistik finde opgørelser om antal personer i de forskellige strata i populationen i året for dataindsamlingen.[21]

Jeg kunne nu vælge at bruge oplysningerne på to forskellige måder. Den ene måde er at medtage strata-variablen direkte i svydesign-komandoen sammen med vægtvariablen:

```
Design3 <- svydesign(
  id = ~1,
  strata = ~pstrata,
  weights = ~V153,
  data = ISSP17)
```

Variablen `pstrata` har 12 koder lavet ud fra en gruppering af vægtvariablen. Fordelen ved at medtage denne er, at man får mindre standardfejl, hvis variansen internt i de enkelte strata er mindre end den samlede varians på den afhængige variabel. Man begår dog også en lille fejl, fordi man med kommandoen foregiver, at der skulle være tale om design-stratifikation og ikke post-stratifikation, og en stikprøve, der post-stratificeres, har lidt større varians på estimaterne end tilsvarende i en design-stratificeret stikprøve (Glasgow, 2005). Det er imidlertid en helt normal praksis, og som Graham Kalton (2021) skriver, vil fejlen være lille, hvis det forventede antal enheder i de enkelte strata ikke er meget lille: "Provided that the expected sample sizes in the poststrata are sufficiently large (say 10 or more), the variance of the poststratified mean is approximately equal to that of a proportionate stratified sample mean based on the same strata." Dette retfærdiggør også den procedure, der følges i eksemplet med ESS-dataene i Afsnit 3. I en vignette til

[20] Downloaded fra https://www.surveybanken.aau.dk/

[21] I Danmark er vi så heldige, at vi ved befolkningsundersøgelser som fx landsdækkende eller kommunale surveys, har rigtig gode og tilgængelige registeroplysninger om fordelinger i populationen, hvilket er en fordel både i designfasen og ifm. bortfaldsanalyse efter indsamlingen. I de fleste andre lande, og også ved mere specielle populationer i danske undersøgelser, er man ikke så heldig. I en del situationer har man imidlertid adgang til de *univariate* fordelinger i populationen, blot ikke kombinerede fordelinger, dvs. man har måske fordelingen på henholdsvis køn og alder, men ikke fordelingen på kombinationer heraf. Man kan godt benytte denne lidt mangelfulde viden til at beregne en poststratificerings-vægt på baggrund af bortfaldsanalyser. Der findes en speciel, iterativ metode til sådanne beregninger, der kaldes for raking, som sikrer de marginale fordelinger på de medtagede variabler. Der findes i `survey`-pakken i R en procedure for at oprette et design, der er baseret på raking-vejning, men som jeg ikke vil komme nærmere ind på her.

survey-pakken lyder det fx også: "Public-use data sets often come with weights that have been adjusted by post-stratification, raking, or calibration. It is standard practice to ignore this fact and treat the weights as if they were sampling weights." (Lumley, 2023B).

Hvis man imidlertid vil specificere stikprøven lidt mere korrekt som poststratificeret, er der også en mulighed for det i R med `survey`-pakken. I denne situation, hvor selve designet på stikprøven er et simpelt tilfældigt udtræk, vil dette blot skulle beskrives på følgende facon:

```
Design2 <- svydesign(id = ~1, data = ISSP17)
```

I selve designet var der hverken klyngeudvælgelse eller stratificeret udvælgelse, så der skal blot nævnes primary sampling unit, som er den enkelte respondent, samt navnet på dataene, ISSP17. Specificeringen af post-stratifikationen følger efter som en tilføjelse eller korrektion herpå, og her skal der angives antal enheder/personer i hvert enkelt strata i populationen. Først skal der genereres en data frame i R, som indeholder strata-koderne samt antal enheder i populationen, sådan som jeg gør nedenfor med navnet `ps.weights`. Navnet på strata-variablen, her `pstrata`, skal have samme navn som i datasættet, da det er linket hertil; og navnet på variablen, der indeholder populationsfrekvenserne skal altid være `Freq`.

```
ps.weights <- data.frame(
  pstrata = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12),
  Freq = c(298592, 312917, 339304, 352411, 371658, 374716,
           400913, 407727, 347897, 343514, 424882, 387922))
```

Fx er der 298.592 personer i populationens strata 1, som er kvinder på 18-25 år, og der er 312.917 mænd i strata 2 med samme aldersspænd. Man skal altså ikke selv generere endelig vægtvariabel, og der kan godt være defineret en designvægt i den indledende `svydesign`-kommando. Man skal blot oplyse om disse populationstal i poststrataene i et tillæg til beskrivelse af stikprøven med kommandoen `postStratify`:

```
Design2ps <- postStratify(
  Design2,
  strata = ~pstrata,
  population = ps.weights)
```

I analyser benyttes nu `Design2ps` i `svy`-kommandoerne. Herunder viser jeg et enkelt eksempel, hvor jeg igen estimerer gennemsnit og standardfejl for variablen `V39`:

```
svymean(~V39, na.rm = TRUE, Design2ps)
```

```
      mean     SE
V39 7.6794 0.0728
```

Gennemsnittet er det samme, som jeg fik i Afsnit 1, hvor designet blot var vægtvariablen, men standardfejlen er nu en anelse mindre.

Til slut skal det imidlertid understreges, at ovenstående eksempel med benyttelse af populationsstrata ikke er helt efter bogen, da jeg selv har skullet rekonstruere Rambølls procedure for generering af vægtvariabel ved bl.a. selv at finde populationstal. Og jeg kan se ud fra fordelingerne i henholdsvis stikprøve og population samt vægt-værdierne i vægt-variablen, at Rambøll og jeg ikke har benyttet 100 pct. de samme opgørelser. Der er

ganske små uoverensstemmelser. Derfor skal dette sidste eksempel på specifikation af en post-stratificeret stikprøve mest ses som nogle principper eller procedurer i R, man kan følge, *hvis* man fra dataudbydernes side får opgivet denne information. I modsat fald vil jeg anbefale at benytte den metode, der blev gennemgået først i dette appendiks, eller alternativt metoden, der er gennemgået i Afsnit 2, dvs. blot at benytte den medfølgende vægt som en designvægt.

Referencer

- Agresti, A. (2018). *Statistical Methods for the Social Sciences*, 5th Edition (Global). Pearson.
- Arel-Bundock, V. (2023). marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests. R package version 0.13.0, https://CRAN.R-project.org/package=marginaleffects.
- Bolen, K.A., Biemer, P.P, Karr, A.F., Tueller, S., & Berzofsky, M.E. (2016). "Are Survey Weights Needed? A Review of Diagnistic Tests in Regressin Analysis". *Annual Review of Statistics and Its Application*, 3: 375-92.
- Cobben, F. (2009). *Nonresponse in sample surveys: methods for analysis and adjustment*. [Thesis, fully internal, Universiteit van Amsterdam.] Statistics Netherlands.
- Glasgow, G. (2005). "Stratified Sampling Types". *Encyclopedia of Social Measurement*, Vel. 3, Elsevier Inc.
- Kalton, G. (2021). "Stratification" in *Introduction to Survey Sampling*. SAGE Research Methods. DOI: https://doi.org/10.4135/9781071909812
- Kaminska, O. (2023). *Guide to Using Weights and Sample Design. Indicators with ESS Data.*
- Lolle, H.L. (2023). *Effekter og prædiktioner fra logistisk regression med R*. AAU working paper: https://vbn.aau.dk/da/persons/henrik-lauridsen-lolle/publications/
- Long, JA (2022). jtools: Analysis and Presentation of Social Scientific Data. R package version 2.2.0, https://cran.r-project.org/package=jtools
- Lumley, T. (2023A). survey: analysis of complex survey samples. R package version 4.2.
- Lumley, T. (2023B). *Making Use of pre-calibrated weights*. Vignette til survey-pakken: https://cran.r-project.org/web/packages/survey/vignettes/precalibrated.pdf