Statistics Norway
Statistisk sentralbyrå

Nina Hagesæther

Notater

Bruk av applikasjonen Struktur

Stabsavdeling/Seksjon for statistiske metoder og standarder

Innhold

1. Innledning

Dette notatet er lagt opp som en håndbok for bruk av applikasjonen Struktur, utviklet av Leiv Solheim, Matz Ivan Faldmo, Jan Sander og Li-Chun Zhang. Notatet baserer seg i stor grad på en tekst av Leiv Solheim tilgjengelig på Q:\Metodekurs\SM03\tekst, kalt "Prediksjon og usikkerhet i S-KJR modeller - prinsipper, metoder, produksjon og eksempler", et upublisert notat av Li-Chun Zhang og en tidligere versjon av hjelpemenyen tilgjengelig i selve applikasjonen, skrevet av Jan Sander. Notatet er forsøkt skrevet så enkelt som mulig. Alle formlene er lagt i Appendiks, slik at bare de som er interessert i teorien trenger å lese dette. Det viktigste er tross alt å vite hvordan man skal bruke Struktur og å kunne tolke resultatene!

Hvis du har forslag til forbedringer av notatet eller applikasjonen, eller det er noe du ikke forstår eller får til, ta gjerne kontakt med supportgruppen.

Faglige og metodemessige problemstillinger:

- Nina Hagesæther (Seksjon for statistiske metoder og standarder, Oslo)
- Li-Chun Zhang (Seksjon for statistiske metoder og standarder, Kongsvinger)
- Leiv Solheim (Seksjon for samferdsels- og reiselivsstatistikk, Kongsvinger)

Spørsmål av teknisk art:

- Jan Sander (Næringsstatistikk IT, Kongsvinger)
- Matz Ivan Faldmo (Næringsstatistikk IT, Kongsvinger)

1.1 Hva er Struktur, og hva kan applikasjonen brukes til?

Applikasjonen startes ved å klikke på *Start*, *Mitt SSB* og dobbeltklikke på ikonet Struktur.

Struktur er en SAS applikasjon som brukes for å estimere totaler og totalenes usikkerhet i utvalgsundersøkelser og for å kjøre statistiske kontroller som grunnlag for revisjon. Den er menybasert og meget enkel å ta i bruk. To SAS datasett kreves for beregningene; en populasjonsfil og en utvalgsfil. Du ledes gjennom fem arkfaner, der ønskede valg gjøres underveis. En hjelpemeny er tilgjengelig gjennom alle arkfanene. Beregningene som utføres av Struktur kan baseres på tre ulike modeller (homogen modell, ratemodell eller enkel regresjonsmodell), der du velger den som passer best til ditt formål.

Struktur kan beregne flere ting, og det er opp til deg hvilke:

- Estimering av ukjente totaler (obligatorisk), på stratum- og landsnivå.
- Variansestimering. Flere ulike robuste variansestimater gis (se senere avsnitt for detaljer).
- Parameterestimater. Ut fra hvilken modell du velger, beregnes parameterestimatene for tilhørende modell.
- Kontroll, en oppsummering av variable innen strata.
- Regresjonsdiagnostikk. Dersom noen av enhetene har stor innflytelse på estimatene, avsløres dette her.
- Vekter eller oppblåsningsfaktorene for den spesifiserte modellen.

1.2 Hvem kan benytte seg av Struktur?

Det er et mål at flest mulig av fagseksjonene skal kunne bruke Struktur. Forløperen til Struktur, S-KJR, er benyttet ved seksjonene 280, 430, 440 og 460, til blant annet følgende statistikker/områder:

- KOSTRA (Kommune-Stat-rapportering)
- Jordbruksstatistikk
- IKT (Informasjons- og kommunikasjonsteknologi)
- Detaljomsetningsindeksen
- Investeringsstatistikk

1.3 Krav til datasett

Før du starter må du ha to SAS-datasett klare, en populasjonsfil og en utvalgsfil. De må inneholde følgende variable:

- Stratumvariable av karakterformat. Gjelder BEGGE datasettene.
- Identifikator av numerisk format. Gjelder BEGGE datasettene.
- En startvekt av numerisk format (hvis du skal beregne vekter). Gjelder BEGGE datasettene.
- En statistikkvariabel av numerisk format. Gjelder KUN utvalgsfilen.
- En forklaringsvariabel av numerisk format. Gjelder BEGGE datasettene.

Det står mer om krav og spesifikasjoner i de påfølgende avsnittene. Eksempler på enkle kommandoer for å konvertere en numerisk variabel til en karaktervariabel og motsatt er gitt i Håndbok i SAS (Lønø, 2000):

```
karakter = COMPRESS(PUT(numerisk, Z2.));
numerisk = karakter + 0;
```

Her er numerisk den numeriske variabelen og karakter er karaktervariabelen. Variabelen foran = er den du oppretter, mens den etter = må finnes på datasettet.

2. Bruk av Struktur

Struktur er en menybasert applikasjon. Gjennom fem arkfaner, Modell, Filer, Variabler, Stratum og Grupper, blir du bedt om å spesifisere valg. For å bevege deg mellom fanene kan du enten klikke på fanenavnet eller knappene *Forrige* / *Neste*. Alle valg gjøres ved markering av opplistede muligheter. *Hjelp*-knappen er tilgjengelig hele tiden, og her kan du få forklaringer og tips slik at du lettere kan gjøre de riktige valgene. *Tilbakeblikk*-knappen gir en oppsummering av valgene du har foretatt, og er også tilgjengelig hele tiden.

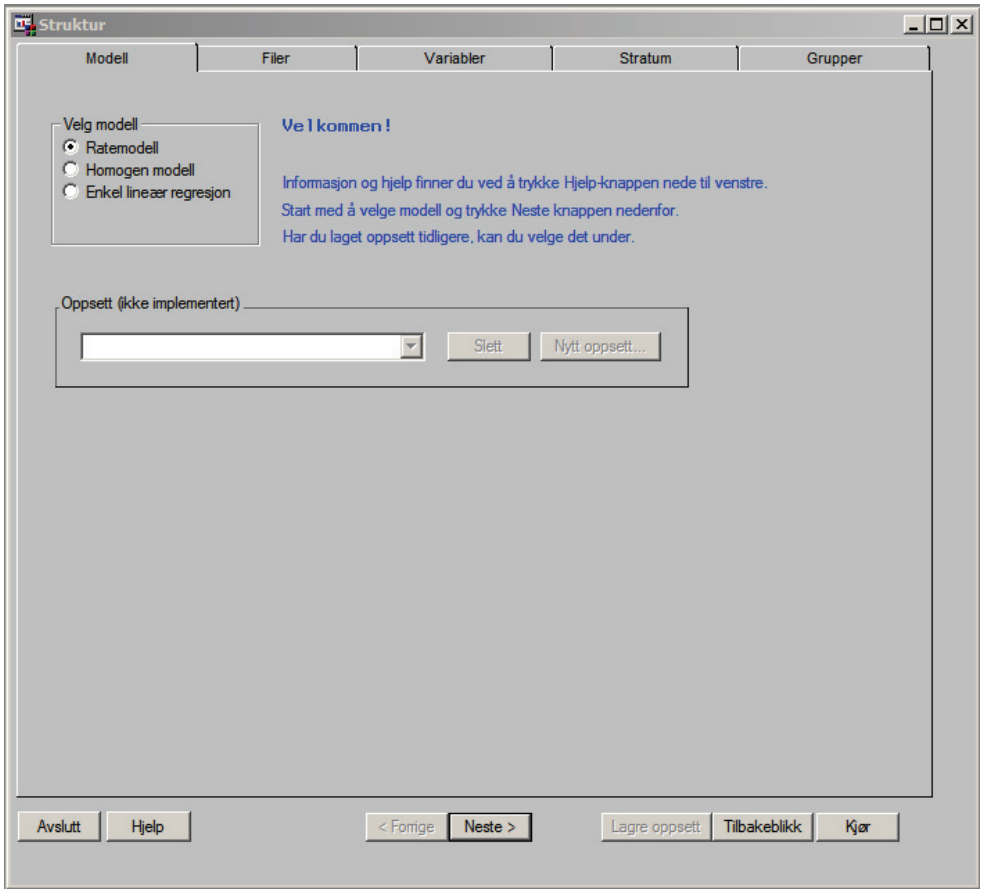
Vi skal nå gå gjennom valgene steg for steg, i den rekkefølgen som er gitt i Struktur. Overskriftene 2.1-2.5 svarer til en arkfane hver.

2.1 Modell

Du får valget mellom tre ulike modeller:

- Homogenmodell
- Ratemodell
- Enkel lineær regresjonsmodell

Under *oppsett*-knappen kan du velge tidligere oppsett (foreløpig ikke implementert). I de tre påfølgende avsnittene skal vi se nærmere på modellene og gjøre rede for hvilke antagelser som ligger til grunn for de enkelte. Generelt ser vi bort fra frafall og andre problemer med data, slik at vi kan bruke hele utvalget til å estimere de ukjente parametrene, den ukjente totalen og beregne usikkerheten til denne.



2.1.1 Homogen modell

Den stratifiserte homogene modellen kan beskrives på følgende måte:

$$y_{hi} = \mu_h + \varepsilon_{hi} \quad ; \quad i = 1,2,...,N_h \quad ; \quad Var(\varepsilon_{hi}) = \sigma_h^2$$

Stratum betegnes med h, statistikkvariabelen med y, enhet med i, gjennomsnitt i populasjonen med μ, feilledd med ε og antall enheter i populasjonen med N. Vi antar at for et gitt stratum bestemmes statistikkvariabelen av et gjennomsnitt som er felles for hele stratumet, pluss et individuelt

feilledd. Vi antar også at alle feilleddene har samme standardavvik innenfor samme stratum. Dette er en modell som brukes veldig ofte i utvalgsundersøkelser der person er enhet.

Den homogene modellen ble for eksempel brukt for å produsere foreløpige tall i Jordbrukstellingen 1999 (se Solheim, Faldmo og Sve, 2002), der det ble stratifisert etter landsdel, jordbruksareal, kornareal og antall kyr. Motivasjonen for modellen var at for et gitt stratum (samme landsdel, jordbruksareal, kornareal og antall kyr) var for eksempel statistikkvariabelen "antall arbeidstimer i året på bruket" ganske lik for alle bruk. Stort sett ga dette svært gode resultater sammenliknet med de endelige tallene.

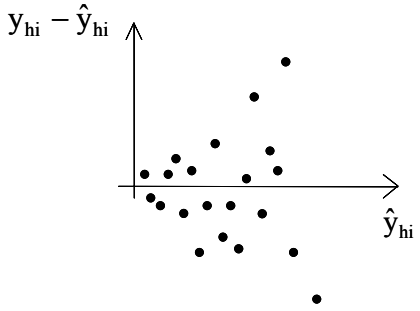
Dersom hver enhet i utvalget tildeles en vekt basert på den homogene modellen, vil summen av vektene gi oss tilbake den totale summen av enheter i populasjonen. Vi kan derfor si at modellen er konsistent med antallet enheter i populasjonen.

2.1.2 Ratemodell

Den stratifiserte ratemodellen kan beskrives på følgende måte:

$$y_{hi} = \beta_h x_{hi} + \varepsilon_{hi} \quad ; \quad i = 1,2,...,N_h \quad ; \quad Var(\varepsilon_{hi}) = x_{hi}\sigma_h^2$$

Her antar vi at det er en variabel x som bidrar til å forklare statistikkvariabelen, og sammenhengen mellom forklaringsvariabelen og statistikkvariabelen er tilnærmet lineær. Et plott av y mot x vil da fortone seg om en tilnærmet rett linje gjennom origo. For et gitt stratum bestemmes statistikkvariabelen av forklaringsvariabelen x_i multiplisert med stigningstallet eller raten β, som er lik for alle enheter i stratumet, pluss et individuelt feilledd. Avviket fra den rette linjen øker med økende x_i. Det betyr at dersom vi plotter residualene ($y_{hi} - \hat{y}_{hi}$) mot predikerte y-verdier ($\hat{y}_{hi}$) vil vi få et plott som minner om dette:



Ratemodellen brukes ofte når forklaringsvariabelen måler antallet personer som produserer eller forbruker den størrelsen som er målt ved statistikkvariabelen.

Dersom hver enhet i utvalget tildeles en vekt basert på ratemodellen, vil summen av $vekt_i$ multiplisert med x_i gi oss tilbake den totale summen av x_i i populasjonen. Det er derfor riktig å si at ratemodellen er konsistent med totalen til forklaringsvariabelen i populasjonen. Dersom forklaringsvariabelen er én for alle personer innen et stratum, blir ratemodellen lik den homogene modellen, og β representerer da gjennomsnittet i stratumet.

2.1.3 Enkel lineær regresjonsmodell

Den enkle lineære regresjonsmodellen kan beskrives på følgende måte:

$$y_{hi} = \alpha_h + \beta_h x_{hi} + \varepsilon_{hi} \quad ; \quad i = 1, 2, ..., N_h \quad ; \quad Var(\varepsilon_{hi}) = \sigma_h^2$$

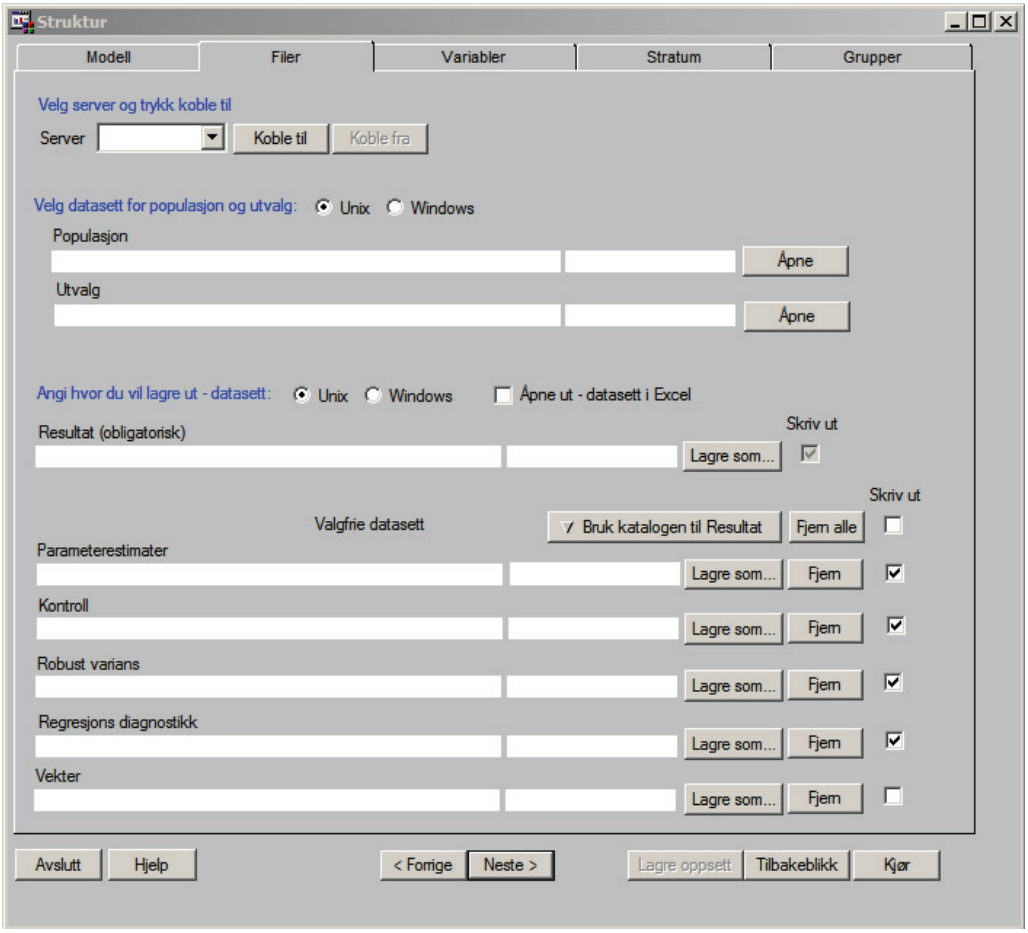
Også her antar vi at det er en variabel x som bidrar til å forklare statistikkvariabelen, og sammenhengen mellom forklaringsvariabelen og statistikkvariabelen er tilnærmet lineær. Et plott av y mot x vil fortone seg om en tilnærmet rett linje som krysser y-aksen i α (i motsetning til 0 som for ratemodellen). Vi antar altså at verdien av statistikkvariabelen bestemmes ved en gjennomsnittlig stratumverdi, en forklaringsvariabel multiplisert med en rate felles for hele stratumet, samt et individuelt feilledd. Avviket fra den estimerte regresjonslinjen antas å være lik for alle statistikkvariabler. I stedet for å få et residualplott som i avsnitt 2.1.2 vil vi få residualer som fordeler seg tilfeldig rundt $\hat{y}_{hi}$.

Denne modellen kan være et naturlig valg dersom vi for utvalget observerer et plott av x mot y som beskrevet over. En annet argument for å velge en lineær regresjonsmodell er at selv om x er 0, antar vi at y kan være ulik 0. Som et eksempel, la x være ansatte og y omsetning for en bedrift. Vi kan velge en lineær regresjonsmodell dersom vi tenker oss at selv om en bedrift ikke har noen ansatte, kan allikevel omsetningen være ulik 0.

En lineær regresjonsmodell er en utvidelse av ratemodellen i det α inkluderes, og av den homogene modellen, i det x inkluderes. Summen av vekt_i multiplisert med x_i vil på samme måte som for ratemodellen gi oss tilbake den totale summen av x_i i populasjonen, og summen av vektene (uten å multiplisere med x_i) vil gi oss totalt antall enheter. Modellen er derfor konsistent med både totalen til forklaringsvariabelen og antall enheter i populasjonen.

2.2 Filer

Denne fanen er uavhengig av hvilken modell du velger.



Beregningene kjøres på Unix, så det første du må gjøre er å logge deg på. Oslo-brukere velger Ovibos eller Ursus, mens Kongsvinger-brukere velger Kodiak eller Sarepta. Det er en fordel å velge den serveren som hjemmekatalogen din ligger på. Trykk *Koble til*-knappen og skriv inn brukernavn og passord. Hvis tilkoblingen lykkes, vil *Koble til* dimmes ut, og *Koble fra*-knappen bli tilgjengelig. Hvis ikke dette skjer, har du kanskje skrevet feil passord. Trykk da *Koble til* på nytt. Deretter gjør du følgende:

- Angi plattformen hvor SAS datasettet ditt er lagret (Unix eller Windows, der Windows kan være for eksempel X-området).
- Velg populasjonsfil og utvalgsfil ved å klikke på *Åpne*-knappen og bla deg frem til rette katalog.
- Velg plattformen du vil lagre resultatene på.
- Angi om du vil åpne resultatene i Excel. Resultatene åpnes da automatisk i Excel etter at kjøringene er fullført.

Etter at filene er på plass angir du hvilke beregninger du ønsker at Struktur skal utføre, ved å klikke på *Lagre som*-knappen. Velg området du vil lagre resultatene på. Et filnavn blir automatisk foreslått, men du kan selvfølgelig endre dette. *Resultat* er obligatorisk, mens de andre beregningene er valgfrie. Ønsker du å utføre alle sammen, kan du enkelt og greit klikke på *Bruk katalogen til Resultat*. Da får alle resultatfilene foreslåtte navn og lagres i samme katalog som *Resultat*. Dersom du har valgt en

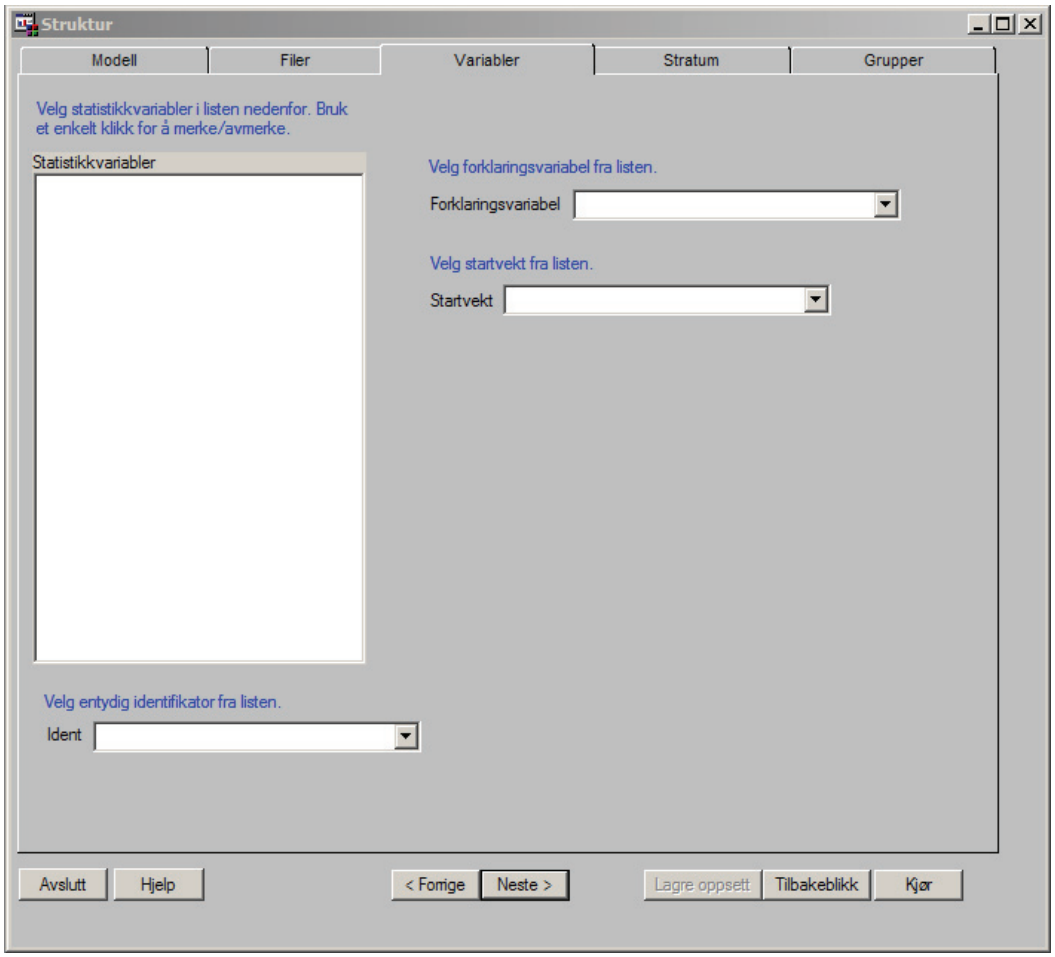
beregning du allikevel ikke ønsker, kan du fjerne den ved å klikke på *Fjern*-knappen. *Skriv ut* kan markeres, og du får da en utskrift av resultatene i SAS-output vinduet. Det kan være hensiktsmessing å ikke skrive ut vekter, ettersom listen gjerne blir lang. Legg også merke til at dersom du velger å beregne vekter, må du angi startvekter på neste arkfane.

Dersom du er interessert i variansestimatet, anbefaler vi at du alltid velger å kjøre robust varians i tillegg til obligatorisk resultat.

Vi skal fortsette å gå gjennom hver arkfane i de påfølgende avsnittene, og kommer tilbake til resultatene i avsnitt 3. Der skal vi se nærmere på hvilke resultater vi bør legge vekt på i forskjellige situasjoner, og dessuten forklare hva resultatene egentlig betyr.

2.3 Variabler

Denne fanen endrer seg etter hvilken modell du velger, og etter om du velger å beregne vekter eller ikke. For alle modeller må du angi en eller flere statistikkvariabler og en identifikator (kalt ident). Velger du ratemodell eller enkel lineær regresjonsmodell må du i tillegg angi en forklaringsvariabel. Velger du å beregne vekter, må du her angi startvekter. Ingen variabelnavn kan ha mer enn 27 tegn. Variabler som finnes både i populasjonsfilen og utvalgsfilen må ha samme variabelnavn i begge datasettene.



2.3.1 Statistikkvariabler

En statistikkvariabel er den avhengige variabelen du ønsker å finne estimater for. Du kan velge flere statistikkvariable ved å markere dem, men maksimalt 50. Har du valgt en variabel du allikevel ikke vil ha med, klikk på den en gang til slik at den ikke lenger er svart. Variabelen må være numerisk og kun finnes på utvalgsfilen (hvis variabelen finnes på populasjonsfilen har vi jo fulltelling, og estimering er ikke nødvendig!).

Fra Arbeidskraftsundersøkelsen (AKU) kan et eksempel på en statistikkvariabel være arbeidsledighetsstatus, der verdien 1 betyr at personen er arbeidsledig og 0 betyr ikke arbeidsledig. Dersom det er KOSTRA-tall som skal analyseres, kan statistikkvariabelen for eksempel være brutto driftsinntekter for en kommune.

2.3.2 Startvekt

Vektene blåser opp tall fra utvalgsnivå til populasjonsnivå. Ofte vil man ta hensyn til trekksannsynlighetene når man beregner vekter, ved å la startvekter være lik invers av trekksannsynligheten. Dersom alle enheter har like startvekter vil vektene for henholdsvis homogen modell, ratemodell og enkel lineær regresjonsmodell bli som beskrevet i Appendiks. Det vil si at dersom du har et enkelt tilfeldig utvalg (alle enheter har lik trekksannsynlighet) spiller det ingen rolle om du lar startvekten være lik invers av trekksannsynlighet eller om du rett og slett lar startvekten være for eksempel 1 for alle enheter. Vektene har egenskaper som nevnt i avsnitt 2.1. Starter du med ulike vekter blir vektene noe annerledes enn beskrevet i Appendiks, men de vil fremdeles ha de nevnte egenskapene.

2.3.3 Ident

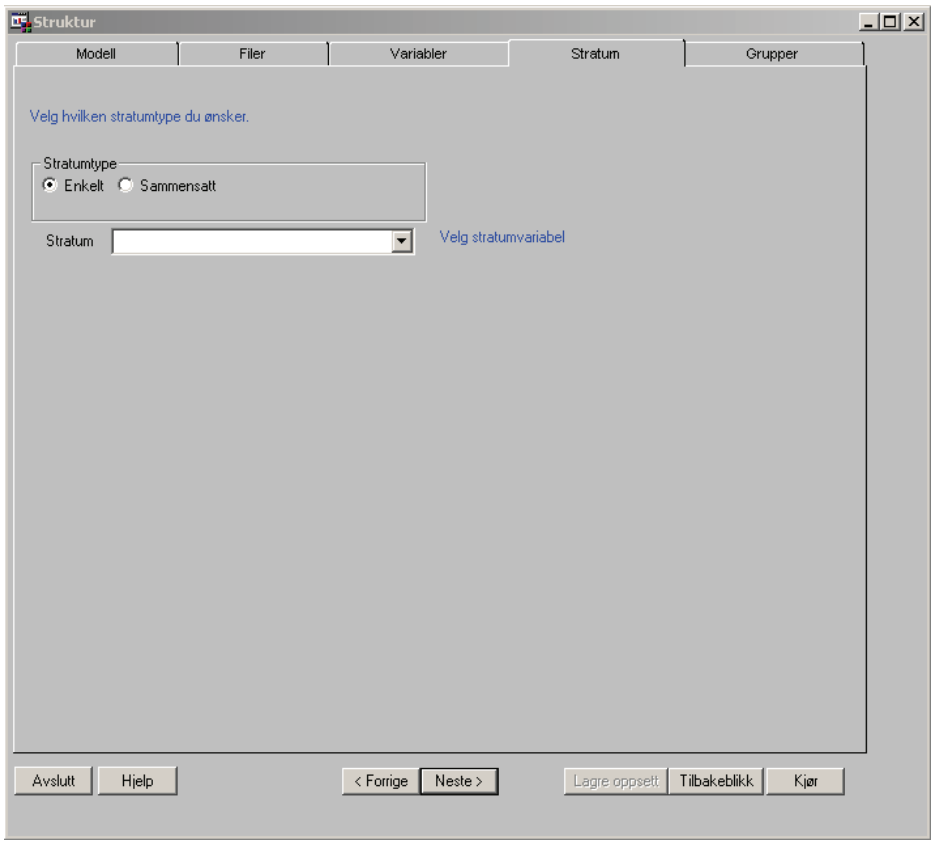
Med en identifikator menes en variabel som angir enhetene, der variabelen er unik for hver enhet. For AKU kan dette være fødselsnummer, siden enhetene er personer. For KOSTRA kan identifikatoren være kommune. Variabelen må være numerisk og finnes på både populasjons- og utvalgsfilen. En identifikator er blant annet nødvendig for å utføre regresjonsdiagnostikken.

2.3.4 Forklaringsvariabel

Forklaringsvariabel må angis om du har valgt rate- eller enkel lineær regresjonsmodell. Det er den variabelen som er med på å forklare variasjonen i statistikkvariabelen. Dersom statistikkvariabelen er omsetning for en bedrift kan forklaringsvariabelen være for eksempel antall ansatte i bedriften. For brutto driftsinntekter rapportert gjennom KOSTRA kan det være antall innbyggere i kommunen. Vi regner altså med at det er en viss sammenheng mellom omsetning og antall ansatte, og mellom driftsinntekter og innbyggere. Forklaringsvariabelen kan ikke inneholde negative tall, den må være numerisk og finnes på både populasjons- og utvalgsfilen.

2.4 Stratum

Her velger du stratum som landet er delt inn i. Stratum er obligatorisk, kan ikke inneholde manglende verdier, må være karaktervariabel og finnes på både populasjons- og utvalgsfilen. Modellen du har valgt tilpasses innen hvert stratum, det vil si at kun enheter fra det gjeldende stratumet vil bli brukt til estimering av verdier for dette stratumet. Du får altså estimeringsresultater for hvert stratum separat, og i tillegg hele landet i ett. Du kan velge mellom to stratumtyper; enkel eller sammensatt. Ønsker du ikke å dele datasettet inn i stratum, kan du velge/opprette en variabel som er lik for alle enheter.



2.4.1 Enkel stratumtype

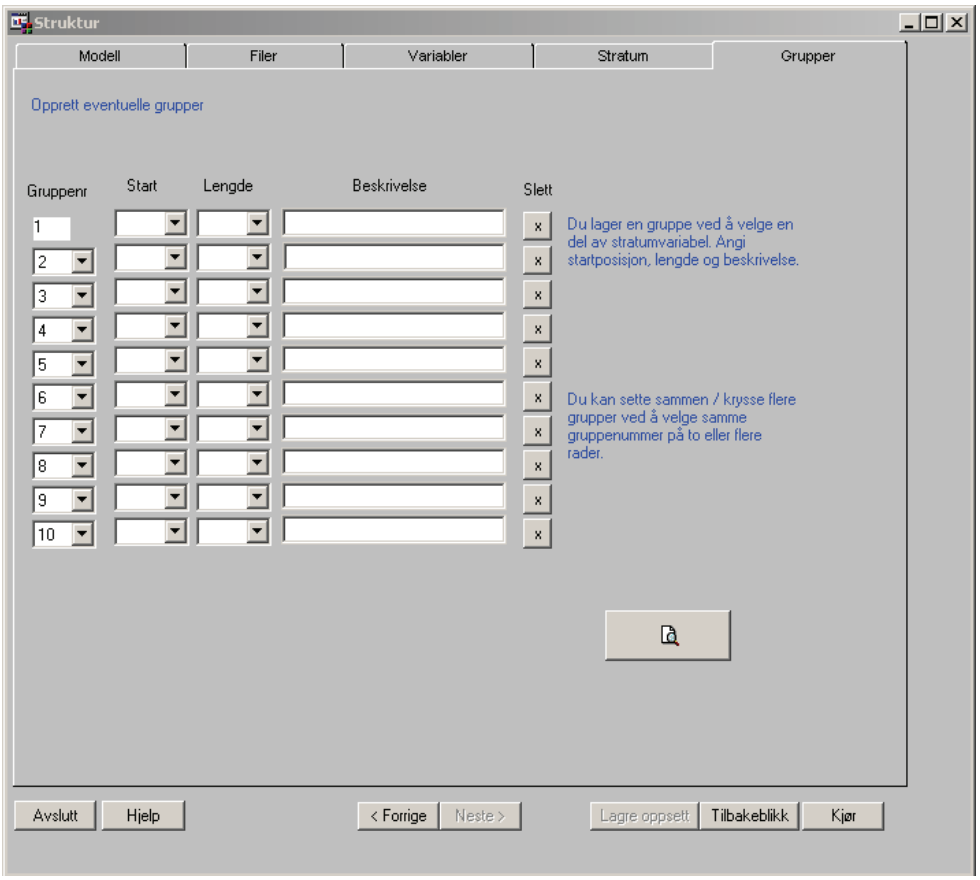
Denne velger du om du allerede har stratumvariabelen på datasettene. For eksempel dersom du ønsker å stratifisere AKU etter alder og kjønn, må det finnes en variabel som kombinerer disse til én variabel på begge datasettene. Dersom kjønn er 1 for menn og 2 for kvinner, og alder går fra 01 (16-19 år) til 12 (70-74 år), kan stratumvariabelen (av typen karaktervariabel) for eksempel se slik ut for en 18 år gammel mann: 101. For en kvinne på 74 år vil stratumverdien være 212. KOSTRA datasettet kan stratifiseres etter SSB-standarden KOSTRA-gruppe, definert ved folkemengde, bundne kostnader per innbygger og frie disponible inntekter per innbygger.

2.4.2 Sammensatt stratumtype

Her kan du konstruere strata ved å sette sammen variable som finnes på begge datasettene. Skjermbildet endrer seg i det du velger sammensatt stratumtype, og variablene du kan velge dukker opp i rullemenyen. Dersom vi igjen ønsker å stratifisere AKU etter alder og kjønn, men ikke har laget stratumvariabelen på forhånd, kan vi hente inn alder og kjønn ved hjelp av rullemenyene. Da blir stratumvariabelen lik som i avsnitt 2.4.1.

2.5 Grupper

Denne fanen avhenger av hva slags stratumtype du velger på fanen Stratum. Ved å opprette grupper kan du aggregere estimater over grupper, der stratum er den mest detaljerte gruppeinndelingen du kan ha. Det vil si at modellen fortsatt er tilpasset innen stratum, men du kan få Struktur til å beregne andre aggregerte estimater enn kun landstotalen (alle strata til sammen). Et eksempel er gitt nedenfor.



Når enkelt stratum er valgt, lager du en gruppe ved å velge en del av stratumvariabelen. Angi startposisjon i stratumvariabelen samt lengde og skriv inn en kort beskrivelse. Du kan vise innholdet i de 10 første gruppevariablene på populasjonen og utvalget ved å trykke på forhåndsvisning-knappen. Dermed kan du kontrollere at du har truffet riktig med posisjonsangivelsen. Som et eksempel, la oss anta at du velger å stratifisere AKU som i avsnitt 2.4.1, altså med tre posisjoner der den første angir kjønn og de to siste alder. Dette gir opphav til 2 kjønn x 12 aldersklasser = 24 strata. Nå vil vi ha resultater for nivåene kjønn og alder separat i tillegg, det vil si at vi også får 2 + 12 = 14 estimater, et for hvert kjønn og et for hver aldersklasse. Da gjør vi følgende:

- For Gruppenr 1, velg 1 for "start" og 1 for "lengde" og kall for eksempel "Beskrivelse" for "Kjønn".
- For Gruppenr 2, velg 2 for "start" og 2 for "lengde", og kall for eksempel "Beskrivelse" for "Alder".

Det er mulig å krysse grupper. Anta for eksempel at i tillegg til kjønn og alder stratifiserer vi AKU etter sysselsetting i registeret. Stratum består nå av tre siffer for kjønn og alder, pluss et siffer som kan være 0 for registersysselsatte personer og 1 for ikke-registersysselsatte personer. Hvis vi ønsker at en gruppe skal bestå av for eksempel kjønn og registerstatus for sysselsetting, gjør vi følgende:

- Lag gruppe 1 med startposisjon 1 og lengde 1
- Lag gruppe 2 med startposisjon 4 og lengde 1
- Velg 1 på rullemenyen under "Gruppenr"

Når sammensatt stratum er valgt, kan du velge enkeltvis som gruppe de variablene som stratum er satt sammen av. Du kan også krysse grupper ved å velge samme gruppenummer som forrige rad, på samme måte som ved enkelt stratum.

Det er beskrivelsen av gruppevariabelen som kommer i utskriften. Det er derfor en fordel at denne er kort og presis. For å angi beskrivelse for en variabel (her stmnace) i SAS kan du bruke denne kommandoen:

```
label stmnace = 'Sysselsetting i registeret';
```

2.6 Klar - ferdig - kjør!

Nå har vi gått gjennom alle valgene, og da kan du kjøre programmet ved å klikke på *Kjør*. Dersom det er noe du har glemt, får du en beskjed og muligheten til å rette på det før du igjen kjører applikasjonen. Kjøringen skal gå raskt, men tiden vil avhenge av antall enheter i datasettene dine.

3. Resultater

I dette avsnittet skal vi se nærmere på resultatene Struktur beregner, og hvilke vi bør legge vekt på i forskjellige situasjoner. Struktur lagrer alle resultatene du ber om som SAS-datasett på det spesifiserte området. Disse datasettene kan brukes videre i statistikkproduksjonen. Selve hovedresultatene (kalt Resultat) blir skrevet ut automatisk til et eget vindu i SAS, andre resultater (kalt Parameterestimater, Kontroll, Robust varians, Regresjonsdiagnostikk og Vekter), kan du velge om skal skrives ut. Vi anbefaler at du alltid velger robust varians dersom du er interessert i usikkerheten til estimatet av totalen. I tillegg kan du velge å åpne resultatene i Excel, noe som kan være en fordel om du skal lage nye tabeller eller bruke resultatene i videre beregninger.

Dersom det er færre enn to enheter i minst ett stratum, får du en melding om du ønsker å fortsette (det er jo umulig å estimere varians for én observasjon!). Velger du ja, vil du ikke få variansestimat for de aktuelle strataene (se for eksempel avsnitt 3.4, KOSTRA gruppe 9 og 15, der 9 har en observasjon og 15 ingen).

3.1 Resultater

Vi starter med et eksempel:

Variabelen av interesse (statistikkvariabelen) er antall arbeidsledige i henhold til Arbeidskraftsundersøkelsen (AKU). Vi kjenner ikke antall arbeidsledige i populasjonen, og ønsker å beregne dette, både for hele landet og i enkelte strata. Vi lar populasjonsfilen være AKU utvalget for første kvartal 2005. Utvalgsfilen konstrueres på følgende måte: Populasjonsfilen sorteres etter fødselsnummer, og de 5 000 med lavest fødselsnummer defineres som utvalget. Identifikator er posisjonen til enheten i den sorterte populasjonsfilen. En homogen modell velges, fordi vi her har med et personutvalg og en 0/1 variabel å gjøre. Dessuten ønsker vi et etterstratifisert estimat (etterstratifiserte estimater brukes i dagens AKU opplegg). Ved å stratifisere og tilpasse en homogen modell innen hvert stratum, oppnår vi et etterstratifisert estimat, der etterstrata er lik de strata vi

tilpasser modellen innen. Sammensatt stratum konstrueres i Struktur av variablene *Sysselsetting i register* (1, 2, 3 eller 4) og *kjønn* (1 eller 2). Grupper er *Sysselsetting i register* og *kjønn*.

Ved kun å velge de obligatoriske resultatene, får vi en firedelt utskrift som vist nedenfor; en for resultater for hele landet, en for resultater per stratum, en for gruppen *Sysselsetting i register* og en for gruppen *kjønn*.

Homogen modell - predikerte totaler, variasjonskoefisienter og konfidensintervall

MODELL	LAND	Sum_N_pop	T_ledig	CV_ledig	LB_ledig	UB_ledig
Homogenmodell	1	21525	612,79	7,09304	527,60	697,98

Homogen modell - predikerte totaler, variasjonskoefisienter og konfidensintervall

MODELL	stratum	Sum_N_pop	T_ledig	CV_ledig	LB_ledig	UB_ledig
Homogenmodell	11	363	4,78	88,91758	-3,55	13,10
Homogenmodell	12	104	3,47	84,35274	-2,26	9,20
Homogenmodell	21	2262	8,92	62,22253	-1,96	19,81
Homogenmodell	22	518	4,21	87,32401	-3,00	11,42
Homogenmodell	31	4700	43,72	27,65526	20,02	67,42
Homogenmodell	32	5945	50,96	25,14122	25,85	76,07
Homogenmodell	41	3532	273,28	10,30319	218,10	328,47
Homogenmodell	42	4101	223,45	11,96803	171,04	275,87

Homogen modell - predikerte totaler, variasjonskoefisienter og konfidensintervall

MODELL	Sysselsetting i register	Sum_N_pop	T_ledig	CV_ledig	LB_ledig	UB_ledig
Homogenmodell	1	467	8,24	62,55447	-1,86	18,35
Homogenmodell	2	2780	13,13	50,70362	0,08	26,19
Homogenmodell	3	10645	94,68	18,60621	60,15	129,21
Homogenmodell	4	7633	496,74	7,81761	420,62	572,85

Homogen modell - predikerte totaler, variasjonskoefisienter og konfidensintervall

MODELL	Kjønn	Sum_N_pop	T_ledig	CV_ledig	LB_ledig	UB_ledig
Homogenmodell	1	10857	330,71	9,50410	269,10	392,31
Homogenmodell	2	10668	282,09	10,64317	223,24	340,93

Følgende estimater listes ut:

- T_ledig: Et estimat for totalt antall arbeidsledige.
- CV_ledig: Et estimat for variasjonskoeffisienten (coefficient of variation = CV).
- LB_ledig og UB_ledig: Nedre og øvre grense (lower bound) for prediksjonsintervallet.

3.1.1 Coefficient of variation

Anta at vi estimerer en total (for eksempel totalt antall arbeidsledige i Norge, som i eksempelet over). Standardavviket til den estimerte totalen forteller oss noe om hvor usikkert estimatet er. Ofte er det slik at jo større estimatet er, jo større standardavvik kan vi akseptere. Det kan derfor være hensiktsmessig å uttrykke standardavviket som en andel av selve estimatet. Estimatet av denne

andelen kaller vi for estimert variasjonskoeffisient eller bare variasjonskoeffisient, og den er oppgitt i prosent i Struktur. Hvor stor prosentandel som er akseptabel kan variere fra statistikk til statistikk. Vi skal her kalle variasjonskoeffisienten for CV.

Av definisjonen over ser vi at CV avhenger av to ting; standardavviket til et estimat og estimatet selv. Jo høyere standardavvik, jo høyere CV, og jo høyere estimat, jo lavere CV. I AKU eksempelet er CV for hele landet på cirka 7 %. Forklaringen på høy CV innen strata ligger i at antall arbeidsledige er veldig lavt, det vil si et lavt estimat.

Dersom du kjører robust variansestimering, lister Struktur ut tre andre estimater for CV. Det er ikke gitt hvilken CV som er den beste i alle tilfeller, men i avsnitt 3.4 kan du få en indikasjon på hvilken som kan være best egnet i ditt tilfelle.

3.1.2 Prediksjonsintervall

Struktur regner ut et 95 % prediksjonsintervall for den sanne totalen, basert på den estimerte totalen. Det er 95 % sjanse for at intervallet inneholder den sanne totalen. Fra AKU eksempelet ser vi at i strata med få arbeidsledige har prediksjonsintervallene en tendens til å inkludere negative tall, noe som selvfølgelig er umulig.

3.2 Parameterestimater

Nå går vi over til Kommune-Stat-Rapportering (KOSTRA) som eksempel.

Målet er å beregne estimater for brutto driftsinntekt for alle kommuner på grunnlag av de kommunene som har rapportert til KOSTRA i 2002. Dersom vi antar at det er en sammenheng mellom brutto driftsinntekt for en kommune (statistikkvariabelen) og kommunens innbyggertall (forklaringsvariabelen), og det er naturlig å i tillegg anta at en kommune med 0 innbyggere også har 0 i brutto driftsinntekt, kan en ratemodell være passende. Stratum er KOSTRA-grupper, nummerert fra 1 til 16. Innen hver KOSTRA gruppe antar vi at raten er lik, det vil si at forholdet mellom brutto driftsinntekter og bosatte i kommunen er omtrent likt for alle kommuner innen et stratum.

Ratemodell - parameterstimater

Obs	kostra_gruppe_02	N_utv	N_pop	X_UTV	X_POP	BETA_bruttoinntekt
1	1	30	46	113692	172551	38.5042
2	2	31	50	93601	156285	44.9601
3	3	17	35	42986	85285	56.7659
4	4	8	11	20321	30381	45.0232
5	5	21	43	49053	94946	50.3320
6	6	22	51	34632	79791	60.9526
7	7	11	13	145904	175198	33.7531
8	8	46	49	630308	674294	34.7019
9	9	1	8	7405	88556	47.2232
10	10	17	27	114538	185506	36.6093
11	11	26	43	230481	346614	39.4746
12	12	5	8	31299	51110	52.2599
13	13	32	36	1234820	1385911	35.9867
14	14	2	3	388122	499129	36.6218
15	16	6	10	6078	9336	97.9446

Det beregnes 15 forskjellige estimater av β, kalt BETA i utskriften (Oslo er fjernet fra datagrunnlaget, fordi kommunen alene utgjør stratum 15, og finnes ikke i utvalget – totalen vil da være et estimat for hele populasjonen med unntak av Oslo). Du kan be om parameterestimater for homogen modell også, da vil utskriften inneholde ALFA. Velger du en regresjonsmodell estimeres både ALFA og BETA.

3.3 Kontroll

Kontroll hjelper deg med å få en oversikt over summen av enheter i populasjonen og utvalget, fordelt på strata. Hvis du har valgt ratemodell eller enkel lineær regresjonsmodell, inneholder utskriften også informasjon om summen av forklaringsvariabelen. Utskriften for parameterestimater inneholder foreløpig den samme informasjonen som Kontroll, men i tillegg parameterestimater for hvert stratum. Det vil si at dersom du har bedt om parameterestimater, trenger du egentlig ikke å be om kontroll i tillegg. For samme eksempel som i avsnitt 3.2 får vi følgende utskrift:

Ratemodell - kontroll av datagrunnlaget

Kostragruppe	N_utv	N_pop	X_UTV	X_POP
1	30	46	113692	172551
2	31	50	93601	156285
3	17	35	42986	85285
4	8	11	20321	30381
5	21	43	49053	94946
6	22	51	34632	79791
7	11	13	145904	175198
8	46	49	630308	674294
9	1	8	7405	88556
10	17	27	114538	185506
11	26	43	230481	346614
12	5	8	31299	51110
13	32	36	1234820	1385911
14	2	3	388122	499129
16	6	10	6078	9336

3.4 Robust variansestimering

Ved å skrive ut Robust varians får du tre nye estimater for CV (se Appendiks for nærmere beskrivelse av estimatorene). Resultatene er vist ved KOSTRA eksempelet.

- CV1: Baserer seg på en variansestimator som består av vektede summer av residualene. Teoretisk sett har CV1 en negativ skjevhet for en gitt utvalgsstørrelse.
- CV2: Som CV1, men residualene er justert i henhold til noe vi kaller h-verdiene. Summen av h-verdiene innen hvert stratum er 1, og enheter med stor verdi for forklaringsvariabelen får stor h-verdi. CV2 er tilnærmet forventningsrett, men kan være noe ustabil dersom det finnes noen veldig store h-verdier. CV2 vil alltid være større enn CV1. CV2 er den som er gjengitt i hovedresultatet som CV.

- CV2A: Som CV2, men justert på en litt annen måte. Regnes for å ha mindre skjevhet enn CV1 og kan være mer stabil enn CV2. Generelt er CV2 tilnærmet lik CV2A, og under den homogene modellen er CV2 = CV2A.
- CV3: Som CV2, men faktoren residualene er justert med er kvadrert. CV3 baserer seg på en konservativ variansestimator som i praksis fungerer som en øvre grense for variansestimatene. Den har en positiv skjevhet, og CV3 > CV2.

Ratemodell - Estimater robust variansestimering

modell	LAND	CV1_bruttoinntekt	CV2_bruttoinntekt	CV2A_bruttoinntekt	CV3_bruttoinntekt
Ratemodell	1	0,26075	0,27499	0,27250	0,29484

Ratemodell - Estimater robust variansestimering

modell	Kostragruppe	CV1_bruttoinntekt	CV2_bruttoinntekt	CV2A_bruttoinntekt	CV3_bruttoinntekt
Ratemodell	1	0,73766	0,75096	0,75082	0,76452
Ratemodell	2	1,16103	1,18559	1,18308	1,21078
Ratemodell	3	2,03774	2,11333	2,11459	2,19254
Ratemodell	4	1,08379	1,16217	1,16028	1,24637
Ratemodell	5	1,09305	1,13676	1,13173	1,18335
Ratemodell	6	3,55727	3,73056	3,66377	3,91365
Ratemodell	7	0,54465	0,57283	0,57443	0,60297
Ratemodell	8	0,34988	0,35423	0,35403	0,35865
Ratemodell	9	0,00000	0,00000	0,00000	0,00000
Ratemodell	10	0,93429	0,96367	0,96538	0,99403
Ratemodell	11	1,03120	1,05158	1,05349	1,07240
Ratemodell	12	2,84089	3,18945	3,18448	3,58133
Ratemodell	13	0,51883	0,53732	0,52907	0,55662
Ratemodell	14	0,39904	0,57759	0,57759	0,85483
Ratemodell	16	5,40900	5,95849	6,01050	6,58563

Så, hvilket estimat av CV bør brukes når? Det finnes ingen fasitsvar, men vi kan få en indikasjon på det beste valget ved å se på differansen mellom de ulike CV ene. Er det liten forskjell, kan valget være bortimot vilkårlig. For KOSTRA eksempelet er differansene små, og alle CV ene vil være et godt valg. CV2 er forventningsrett, så det er fornuftig å velge denne. Er forskjellen mellom CV fra hovedresultatet og de mer robuste CV ene derimot store, bør vi velge en av de mer robuste.

Er differansen mellom CV1 og CV2 stor, er det grunn til å tro at variansestimatet i førstnevnte inneholder en betydelig skjevhet. Er differansen mellom CV2 og CV2A stort, kan førstnevnte påvirkes av unormale observasjoner (disse observasjonene avsløres i neste avsnitt, der vi også forklarer hva som menes med unormal), slik at fordelen med å få et forventningsrett estimat (ved å gå fra CV1 til CV2) går på bekostning av estimatets stabilitet. Da kan det være en fordel å velge for eksempel CV2A. Er alle estimatene varierende, kan det være tryggest å velge det konservative estimatet CV3. Hva som menes med uttrykk som "stor" eller "varierende" er heller ikke entydig, og må vurderes i hvert enkelt eksempel.

3.5 Regresjonsdiagnostikk

I mange praktiske situasjoner vil vi ofte støte på problemet at en eller flere av verdiene til statistikkvariabelen avviker svært mye fra de andre. Vi skal nå se på hvordan vi kan undersøke om en enhet er avvikende i forhold til de andre i utvalget. Selve ideen er enkel, nemlig å utforme kriterier for

når en enhet er farlig i den forstand at den både avviker sterkt fra resten av utvalget og dessuten påvirker resultatet betydelig i tillegg.

KOSTRA eksempelet brukes for vise hvordan resultatene kan se ut dersom du ber om regresjonsdiagnostikk.

Ratemodell - Regresjons diagnostikk

Obs	DIAGNOSTIKK	VERDI	KRITISK_VARIABEL	kommune	kostra_gruppe_02	N_utv	VERDI_Y
1	H	0.13133	bruttoinntekt	1535 Vestnes	5	21	344357
2	H	0.12491	bruttoinntekt	0536 Søndre Land	5	21	306473
3	H	0.11772	bruttoinntekt	1438 Bremanger	6	22	196981
4	H	0.08303	bruttoinntekt	0219 Bærum	13	32	4249177
5	R	5.69166	bruttoinntekt	0941 Bykle	16	6	128036
6	R	5.15247	bruttoinntekt	0219 Bærum	13	32	4249177
7	R	4.88975	bruttoinntekt	1911 Kvæfjord	6	22	266168
8	R	2.91489	bruttoinntekt	1102 Sandnes	13	32	1716887
9	R	2.86701	bruttoinntekt	0826 Tinn	12	5	388775
10	R	2.56850	bruttoinntekt	0819 Nome	11	26	325237
11	R	2.49232	bruttoinntekt	1630 Åfjord	1	30	150141
12	R	2.47192	bruttoinntekt	0540 Sør-Aurdal	1	30	149579
13	R	2.36154	bruttoinntekt	0719 Andebu	1	30	163670
14	R	2.35920	bruttoinntekt	1548 Fræna	7	11	335691
15	R	2.29501	bruttoinntekt	2015 Hasvik	3	17	86958
16	R	2.21482	bruttoinntekt	1503 Kristiansund	8	46	694694
17	R	2.17264	bruttoinntekt	1438 Bremanger	6	22	196981
18	R	2.01942	bruttoinntekt	1824 Vefsn	8	46	549453
19	G	2.29185	bruttoinntekt	0941 Bykle	16	6	128036
20	G	1.55046	bruttoinntekt	0219 Bærum	13	32	4249177

I regresjonssammenheng er en observasjon ekstrem dersom absoluttverdien av tilhørende residual er uvanlig stor, og/eller hvis x_i har en uvanlig plassering blant alle x-ene. Ekstreme observasjoner på den ene eller andre måten trenger nødvendigvis ikke å bekymre oss. En observasjon som er ekstrem på begge måter samtidig kan derimot ha store innvirkninger på resultatene. Vi kaller observasjonene for en kritisk verdi. Kun de tre første med G-verdi større enn den kritiske verdien er tatt med på utskriften (det var i alt 21).

Struktur lister ut tre verdier som kan brukes til diagnostikk (se også Appendiks);

- H er det samme som h-verdien, forklart i ballpunkt 2 i avsnitt 3.4, og er en indikasjon på om observasjonen har en uvanlig x-verdi. Dersom H er større enn $2p/n$ for en observasjon, der p er antall parametere i modellen (en for den homogene modellen og ratemodellen og to for regresjonsmodellen), er x-verdien uvanlig. For den homogene modellen har vi at $x = 1$ og $\mathrm{H} = 1/n$ for alle observasjoner, og dermed overstiger ingen H den kritiske grensen.
- R er absoluttverdien til det jackknife standardiserte residualet, og gir en indikasjon på om residualet til en observasjon er uvanlig stort. R har 2 som kritisk grense, fordi vi antar at de standardiserte residualene er normalfordelte. Dersom R er større enn 2, sier vi at observasjonen har et uvanlig stort residual.
- G er absoluttverdien til DFFITS, og er en standard diagnostikk på mulige kritiske observasjoner for estimatene av regresjonskoeffisientene (α for en homogen modell, β for en

ratemodell og begge for en regresjonsmodell). G er en kombinasjon av R og H, og dens kritiske grense er gitt ved $2(p/n)^{0,5}$.

Alle verdiene kan beregnes i prosedyren *proc reg* i SAS, der de betegnes med henholdsvis h, r og dffits. Den kritiske grensen er basert på en del antagelser, og vi kan derfor være litt fleksible når vi avgjør hvilke observasjoner som er kritiske og ikke.

3.6 Vekter

For å kunne beregne vekter ut fra de forskjellige modellene, må du angi en startvekt. Ønsker du ikke å ta hensyn til trekksannsynligheter, kan startvekten være lik for alle enheter i utvalget. Vektene oppfyller som nevnt tidligere følgende krav:

- Summen av vektene innen et stratum svarer til antall enheter i populasjonsstratumet (homogen modell og enkel lineær regresjonsmodell)
- Summen av vektene multiplisert med forklaringsvariabelen svarer til summen av forklaringsvariabelen i populasjonsstratumet (ratemodell og enkel lineær regresjonsmodell).

Det er teoretisk mulig å få negative vekter. Hvis dette er tilfellet, burde vektene i de aktuelle strataene justeres.

4. Eksempel: Forskning og utvikling 2004

FoU (forskning og utvikling) statistikken går blant annet ut på å kartlegge foretakenes totale utgifter til egenutført FoU (oppgitt i kroner) og FoU-personale i alt (oppgitt i antall personer). Vi skal nå se på hvordan Struktur kan brukes til dette formålet. Vi benytter tall fra 2004.

4.1 Populasjonen og utvalget

Populasjonen består av 11 647 foretak, der kun foretak som har 10 eller flere sysselsatte er inkludert. Foretakene er fordelt på fem sysselsettingsgrupper (10-19, 20-49, 50-99, 100-249 og 250 eller flere sysselsatte) og 41 næringer.

Foretak som har 50 eller flere sysselsatte fulltelles, med unntak av foretak i næring 45 og 51. Disse fulltelles om antall sysselsatte er 100 eller flere. Fulltelte foretak er ikke med i estimeringsopplegget, de får vekt = 1 og teller bare for seg selv. Det finnes også et tilleggsutvalg av foretak som året før var med i utvalget og hadde utgifter til egenutført FoU på 1 000 000 kroner eller mer, og/eller utgifter til innkjøpt FoU på 3 000 000 kroner eller mer. Disse får også vekt = 1 og holdes utenfor estimeringsopplegget. Fulltelling og tilleggsutvalg utgjør til sammen 1 909 foretak, slik at populasjonen det skal trekkes et utvalg (som kalles sannsynlighetsutvalg) fra utgjør 9 738 foretak.

Det endelige utvalget består altså av tre deler; fulltelling, tilleggsutvalg og sannsynlighetsutvalg. De to førstnevnte utgjør til sammen 1 881 foretak, mens sannsynlighetsutvalget utgjør 2 774.

4.2 Estimeringsopplegget

Foretakene i sannsynlighetsutvalget må blåses opp slik at de gjelder for alle 9 738 foretakene i sannsynlighetspopulasjonen. For statistikkvariabelen utgifter til egenutført FoU (INTFOU) brukes en ratemodell som vist i avsnitt 2.1.2, der antall sysselsatte er forklaringsvariabelen. For FoU-personale i alt (FOUPER) brukes en homogenmodell som vist i avsnitt 2.1.1. Identiteten er organisasjonsnummeret.

Det stratifiseres etter næring og sysselsettingsgruppe, til sammen 81 strata. Ved ratemodellen og homogenmodellen kan vi beregne vekter, slik at vi enkelt kan estimere totaler for andre statistikkvariable enn de to vi skal se på her, gitt at også disse statistikkvariablene følger en av de to modellene.

4.3 Resultater

I Struktur får vi resultater per stratum i tillegg til en total. $\hat{T}$ betegner estimatet av totalen mens CV betegner variasjonskoeffisienten. CV en som er oppgitt her, baserer seg *kun* på resultatene fra sannsynlighetsutvalget og sannsynlighetspopulasjonen. I dette estimeringsopplegget har vi i tillegg en fulltelling og et tilleggsutvalg. Her er $Var(\hat{T}-T)=CV(\hat{T}-T)=0$, mens $\hat{T}=T\neq 0$. For å finne den reelle CV en, må vi først beregne variansen basert på tallene i Struktur ved følgende formel:

$$Var(\hat{T}-T)=\sum_h Var(\hat{T}_h-T_h)=\sum_h [CV(\hat{T}_h-T_h)\cdot\hat{T}_h]^2=[CV(\hat{T}-T)\cdot\hat{T}]^2$$

Deretter må vi legge sammen T fra fulltellingen og tilleggsutvalget og $\hat{T}$ fra sannsynlighetsutvalget, betegn denne for $\hat{T}^*$. Nå kan vi beregne CV basert på hele utvalget:

$$CV(\hat{T}-T)=\frac{SE(\hat{T}-T)}{\hat{T}^*}=\frac{[Var(\hat{T}-T)]^{0,5}}{\hat{T}^*}$$

Her ser vi en liten del av resultatene fra Struktur, statistikkvariabelen er INTFOU oppgitt i 1000 kroner (CV er lik dersom alle enhetene skaleres opp eller ned med den samme faktoren). Den øverste tabellen gjelder for hele landet, mens den nederste er et utvalg av strata. I stratum 112 er det fulltelling, og dermed er $CV = 0$. Legg merke til at CV basert kun på sannsynlighetsutvalget er 7,06. Når tall fra fulltelling og tilleggsutvalg er lagt til, blir CV = 1,15.

LAND	Sum_N_pop	Sum_X_pop	T_intfou	CV_intfou
1	9723	221024	2109271,94	7,05865

strata	Sum_N_pop	Sum_X_pop	T_intfou	CV_intfou
051	46	657	35264,58	32,14869
052	18	548	7578,88	27,20116
111	21	292	6258,78	13,45836
112	15	397	7688,00	0,00000
141	46	620	2368,82	26,59498

For hver enhet i utvalget får vi en vekt. Tabellen nedenfor viser vektene ved ratemodellen for et utvalg av strata, og disse er uavhengig av statistikkvariabelen. Alle enhetene i samme stratum får samme vekt. Dersom vi antar at en annen variabel også kan beskrives ved ratemodellen, kan vi enkelt beregne totalen for den nye variabelen ved å multiplisere vekten og verdien for statistikkvariabelen, og så summere over alle enhetene i utvalget.

strata	VEKT_2
51	2,09236
52	1,1939
111	1,03915
112	1
141	1,30802

Appendiks

A.1 Resultater og Parameterestimater

I dette avsnittet skal vi se på utledningene av de verdiene som omtales i avsnitt 3.1 og 3.2, nemlig et estimat av totalen for statistikkvariabelen, prediksjonsvariansen til estimatet av totalen, øvre og nedre grense for prediksjonsintervallet og parameterestimatene.

A.1.1 Homogen modell

Anta at statistikkvariabelen i populasjonen kan beskrives ved en homogen modell. Modellen er gitt ved uttrykket:

$$y_{hi} = \mu_h + \varepsilon_{hi} \quad ; \quad i = 1,2,...,N_h \quad ; \quad Var(\varepsilon_{hi}) = \sigma_h^2$$

Stratum betegnes med h, verdier av statistikkvariabelen med y, enhet med i, gjennomsnitt i populasjonen med μ, feilledd med ε og antall enheter i populasjonen med N. Det er to parametere som må estimeres fra enhetene i utvalget, μ_h og σ_h^2. Estimeringen bygger på minste kvadraters metode, og vi finner følgende estimatorer:

$$\hat{\mu}_h = \frac{1}{n_h}\sum_{i\in s_h} y_{hi} = \bar{y}_{s_h} \tag{1.1}$$

$$\hat{\sigma}_h^2 = \frac{\sum_{i\in s_h}(y_{hi} - \hat{\mu}_h)^2}{n_h - 1} \tag{1.2}$$

Antall enheter i utvalget er gitt ved n_h, og utvalget betegnes s. For å finne et estimat av den ukjente totalen i stratumet, må vi predikere en verdi for alle enhetene utenfor utvalget (enhetene i utvalget kjenner vi jo!). Dette gjør vi ved å sette inn utvalgsgjennomsnittet i (1.1) for hver av verdiene utenfor utvalget: $\hat{y}_{hi} = \hat{\mu}_h$, hvis $i \notin s_h$. Da er estimatoren for totalen gitt ved følgende uttrykk:

$$\hat{T}_h = \sum_{i\in s_h} y_{hi} + \sum_{i\notin s_h} \hat{y}_{hi} = \sum_{i\in s_h} y_{hi} + (N_h - n_h)\hat{\mu}_h = \sum_{i\in s_h} \frac{N_h}{n_h} y_{hi} = N_h \cdot \hat{\mu}_h \tag{1.3}$$

De to siste uttrykkene i (1.3) er hensiktsmessige måter å beregne totalen direkte på. I det nest siste uttrykket er totalen gitt ved å summere over enhetene i utvalget multiplisert med en vekt. Vekten er lik forholdet mellom antall enheter i populasjonen og antall enheter i utvalg og kalles w_{hi}. Summen av vektene vil gi oss tilbake antall enheter i populasjonen:

$$\sum_{i\in s_h} w_{hi} = \sum_{i\in s_h} \frac{N_h}{n_h} = n_h \frac{N_h}{n_h} = N_h \tag{1.4}$$

Modellen er derfor konsistent med antall enheter i populasjonen. For å beregne usikkerheten i prediksjonen i (1.3) kan vi se på det andre uttrykket:

$$V(\hat{T}_h - T_h) = V[(N_h - n_h)\hat{\mu}_h - \sum_{i \notin s_h} y_{hi}] = (N_h - n_h)^2 \frac{\sigma_h^2}{n_h} + (N_h - n_h)\sigma_h^2 = N_h^2 \frac{N_h - n_h}{N_h} \frac{\sigma_h^2}{n_h}$$

Ved å sette inn (1.2) får vi et uttrykk for den empiriske variansen til avviket mellom den predikerte verdien for totalen og totalen selv:

$$\hat{V}(\hat{T}_h - T_h) = N_h^2 \frac{N_h - n_h}{N_h} \frac{\hat{\sigma}_h^2}{n_h} \tag{1.5}$$

Nå kan vi skrive opp standardfeilen (SE), en estimator for variasjonskoeffisienten (CV) og et 95 % prediksjonsintervall (PI) for den ukjente totalen, basert på (1.5):

$$SE(\hat{T}_h - T_h) = N_h \sqrt{\frac{N_h - n_h}{N_h}} \frac{\hat{\sigma}_h}{\sqrt{n_h}}$$

$$CV(\hat{T}_h - T_h) = \frac{SE(\hat{T}_h - T_h)}{\hat{T}_h} = \sqrt{\frac{N_h - n_h}{N_h}} \frac{\hat{\sigma}_h}{\hat{\mu}_h \sqrt{n_h}} \tag{1.6}$$

$$PI = [\hat{T}_h - 1.96 \cdot SE(\hat{T}_h - T_h), \hat{T}_h + 1.96 \cdot SE(\hat{T}_h - T_h)] \tag{1.7}$$

De estimatene som finnes i Resultater i avsnitt 3.1 og Parameterestimater i avsnitt 3.2 er nå gitt ved (1.1), (1.3) og (1.7). Resultatene for Designvekter er gitt ved w_{hi} i (1.4). I stedet for (1.6) har vi valgt å gi en CV som er basert på et robust variansestimat i hovedresultatet, se avsnitt A.4.

A.1.2 Ratemodell

Anta at statistikkvariabelen i populasjonen kan beskrives ved en ratemodell. Modellen er gitt ved:

$$y_{hi} = \beta_h x_{hi} + \varepsilon_{hi} \quad ; \quad i = 1, 2, ..., N_h \quad ; \quad Var(\varepsilon_{hi}) = x_{hi}\sigma_h^2$$

Her er x en kjent forklaringsvariabel og β stingstallet eller raten, som er lik for alle enheter i stratumet. Se ellers avsnitt A.1 for definisjoner. Det er to parametere som må estimeres fra enhetene i utvalget, β_h og σ_h^2. Estimeringen bygger på minste kvadraters metode, og vi finner følgende estimatorer:

$$\hat{\beta}_h = \frac{\sum_{i \in s_h} y_{hi}}{\sum_{i \in s_h} x_{hi}} = \frac{y_{s_h}}{x_{s_h}} \tag{2.1}$$

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} \frac{(y_{hi} - \hat{\beta}_h x_{hi})^2}{x_{hi}} \tag{2.2}$$

Neste trinn er å predikere verdiene til enhetene utenfor utvalget, og det gjør vi ved å multiplisere den estimerte raten med forklaringsvariabelen: $\hat{y}_{hi} = \hat{\beta}_h x_{hi}$, hvis $i \notin s_h$. Da er en estimator for totalen gitt ved følgende uttrykk, der X_h er summen av statistikkvariabelen i populasjonen:

$$\hat{T}_h = \sum_{i \in s_h} y_{hi} + \sum_{i \notin s_h} \hat{y}_{hi} = \sum_{i \in s_h} y_{hi} + (X_h - x_{s_h})\hat{\beta}_h = \sum_{i \in s_h} \frac{X_h}{x_{s_h}} y_{hi} = X_h \cdot \hat{\beta}_h \qquad (2.3)$$

De to siste uttrykkene i (2.3) er hensiktsmessige måter å beregne totalen direkte på. I det nest siste uttrykket er totalen gitt ved å summere over enhetene i utvalget multiplisert med en vekt. Vekten er lik forholdet mellom totalen av x-ene i populasjonen og x-ene i utvalget og kalles w_{hi}. Summen av vektene multiplisert med x vil gi oss tilbake totalen av x-ene i populasjonen:

$$\sum_{i \in s_h} x_{hi} w_{hi} = \sum_{i \in s_h} \frac{X_h}{x_{s_h}} x_{hi} = \frac{X_h}{x_{s_h}} x_{s_h} = X_h \qquad (2.4)$$

Modellen er derfor konsistent med totalen til forklaringsvariabelen i populasjonen. For å beregne usikkerheten i prediksjonen i (2.3) kan vi se på det andre uttrykket:

$$V(\hat{T}_h - T_h) = V[(X_h - x_{s_h})\hat{\beta}_h - \sum_{i \notin s_h} y_{hi}] = (X_h - x_{s_h})^2 \frac{\sigma_h^2}{x_{s_h}} + (X_h - x_{s_h})\sigma_h^2 = X_h^2 \frac{X_h - x_{s_h}}{X_h} \frac{\sigma_h^2}{x_{s_h}}$$

Ved å sette inn (2.2) får vi et uttrykk for den empiriske variansen til avviket mellom den predikerte verdien for totalen og totalen selv:

$$\hat{V}(\hat{T}_h - T_h) = X_h^2 \frac{X_h - x_{s_h}}{X_h} \frac{\hat{\sigma}_h^2}{x_{s_h}} \qquad (2.5)$$

Nå kan vi skrive opp standardfeilen (SE), en estimator for variasjonskoeffisienten (CV) og et 95 % prediksjonsintervall (PI) for den ukjente totalen, basert på (2.4):

$$SE(\hat{T}_h - T_h) = X_h \sqrt{\frac{X_h - x_{s_h}}{X_h}} \frac{\hat{\sigma}_h}{\sqrt{x_{s_h}}}$$

$$CV(\hat{T}_h - T_h) = \frac{SE(\hat{T}_h - T_h)}{\hat{T}_h} = \sqrt{\frac{X_h - x_{s_h}}{X_h}} \frac{\hat{\sigma}_h}{\hat{\beta}_h \sqrt{x_{s_h}}} \qquad (2.6)$$

$$PI = [\hat{T}_h - 1.96 \cdot SE(\hat{T}_h - T_h), \hat{T}_h + 1.96 \cdot SE(\hat{T}_h - T_h)] \qquad (2.7)$$

De estimatene som beregnes for Resultater i avsnitt 3.1 og Parameterestimater i avsnitt 3.2 er nå gitt ved (2.1), (2.3) og (2.7). Resultatene for Designvekter er gitt ved w_{hi} i (2.4). I stedet for (2.6) har vi valgt å gi en CV som er basert på et robust variansestimat i hovedresultatet, se avsnitt A.4.

A.1.3 Enkel lineær regresjonsmodell

For en enkel lineær regresjonsmodell er utregningen noe mer komplisert, men samme fremgangsmåte brukes. En enkel lineær regresjonsmodell kan beskrives på følgende måte:

$$y_{hi} = \alpha_h + \beta_h x_{hi} + \varepsilon_{hi} \quad ; \quad i = 1,2,...,N_h \quad ; \quad Var(\varepsilon_{hi}) = \sigma_h^2$$

Se avsnitt A.1 og A.2 for definisjoner. Det er tre parametere som må estimeres fra enhetene i utvalget, og utregningene bygger på minste kvadraters metode:

$$\hat{\alpha}_h = \bar{y}_{s_h} - \hat{\beta}_h \bar{x}_{s_h} \tag{3.1}$$

$$\hat{\beta}_h = \frac{\sum_{i \in s_h} y_{hi}(x_{hi} - \bar{x}_{s_h})}{\sum_{i \in s_h} (x_{hi} - \bar{x}_{s_h})^2} \tag{3.2}$$

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 2} \sum_{i \in s_h} (y_{hi} - \hat{\alpha}_h - \hat{\beta}_h x_{hi})^2 \tag{3.3}$$

Neste trinn er å predikere verdiene til enhetene utenfor utvalget, og det gjør vi ved følgende uttrykk: $\hat{y}_{hi} = \hat{\alpha}_h + \hat{\beta}_h x_{hi}$, hvis $i \notin s_h$. Da er en estimator for den ukjente totalen gitt ved følgende uttrykk:

$$\begin{aligned} \hat{T}_h &= \sum_{i \in s_h} y_{hi} + \sum_{i \notin s_h} \hat{y}_{hi} = \sum_{i \in s_h} y_{hi} + (N_h - n_h)\hat{\alpha}_h + (X_h - x_{s_h})\hat{\beta}_h \\ &= N_h(\hat{\alpha}_h + \bar{X}_h \hat{\beta}_h) = \sum_{i \in s_h} \{\frac{N_h}{n_h}[1 - (x_{hi} - \bar{x}_{s_h})(\bar{x}_{s_h} - \bar{X}_h)v_{s_h}^{-2}]\} y_{hi} \end{aligned} \tag{3.4}$$

Her er $v_{s_h}^2 = \frac{\sum_{i \in s_h} (x_{hi} - \bar{x}_{s_h})^2}{n_h}$. De to siste uttrykkene i (3.4) er hensiktsmessige måter å beregne totalen direkte på. I det siste uttrykket er totalen gitt ved å summere over enhetene i utvalget multiplisert med en vekt. Summen av vektene vil gi oss tilbake antall enheter i populasjonen, mens summen av vektene multiplisert med x vil gi oss tilbake totalen av x-ene i populasjonen, med samme fremgangsmåte som i avsnitt A.1 og A.2. Modellen er derfor konsistent både med totalt antall enheter i populasjonen og totalen til forklaringsvariabelen. For å beregne usikkerheten i prediksjonen i (2.3) kan vi se på det andre uttrykket:

$$V(\hat{T}_h - T_h) = V[(N_h - n_h)\hat{\alpha}_h + (X_h - x_{s_h})\hat{\beta}_h - \sum_{i \notin s_h} y_{hi}] = N_h^2 [\frac{N_h - n_h}{N_h} + \frac{(\bar{X}_h - \bar{x}_{s_h})^2}{v_{s_h}^2}] \frac{\sigma_h^2}{n_h}$$

Ved å sette inn (3.3) får vi et uttrykk for den empiriske variansen til avviket mellom den predikerte verdien for totalen og totalen selv:

$$\hat{V}(\hat{T}_h - T_h) = N_h^2 [\frac{N_h - n_h}{N_h} + \frac{(\bar{X}_h - \bar{x}_{s_h})^2}{v_{s_h}^2}] \frac{\hat{\sigma}_h^2}{n_h} \tag{3.5}$$

Nå kan vi skrive opp standardfeilen (SE), en estimator for variasjonskoeffisienten (CV) og et 95 % prediksjonsintervall (PI) basert på (3.5):

$$SE(\hat{T}_h - T_h) = N_h \sqrt{\frac{N_h - n_h}{N_h} + \frac{(\bar{X}_h - \bar{x}_{s_h})^2}{v_{s_h}^2}} \frac{\hat{\sigma}_h}{\sqrt{n_h}}$$

$$CV(\hat{T}_h - T_h) = \frac{SE(\hat{T}_h - T_h)}{\hat{T}_h} = \sqrt{\frac{N_h - n_h}{N_h} + \frac{(\bar{X}_h - \bar{x}_{s_h})^2}{v_{s_h}^2}} \frac{\hat{\sigma}_h}{(\hat{\alpha}_h + \hat{\beta}_h \bar{X}_h)\sqrt{n_h}} \tag{3.6}$$

$$PI = [\hat{T}_h - 1.96 \cdot SE(\hat{T}_h - T_h), \hat{T}_h + 1.96 \cdot SE(\hat{T}_h - T_h)] \tag{3.7}$$

De estimatene som beregnes for Resultater i avsnitt 3.1 og Parameterestimater i avsnitt 3.2 er nå gitt ved (3.1), (3.2), (3.4), (3.5) og (3.7). Resultatene for Designvekter er gitt ved {} i (3.4). I stedet for (3.6) har vi valgt å gi en CV som er basert på et robust variansestimat i hovedresultatet, se avsnitt A.4.

A.2 Robust variansestimering

I A.1 er variansstrukturen gitt ved $Var(\varepsilon_{hi}) = \sigma_h^2 v_i$, der vi antar at $v_i = 1$ for den homogene modellen og den enkle lineære regresjonsmodellen, og x_i for ratemodellen. Vi skal nå se på en robust variansestimering for et gitt stratum, og dropper derfor betegnelsen h. Med robust mener vi at i stedet for å anta en variansstruktur v_i, lar vi $Var(\varepsilon_i) = \sigma_i^2$. Usikkerheten i prediksjonen kan deles opp i to; $V(\hat{T}_s - T) = V(\sum_{i \notin s} \hat{y}_i) + V(\sum_{i \notin s} y_i) = V(\hat{T}_r) + V(T_r)$, der r indikerer at summene gjelder for enheter utenfor utvalget.

Det første leddet kan skrives som $V(\hat{T}_r) = \sum_{i \in s} a_i^2 \sigma_i^2$. Vektene a_i er modellavhengige, generelt gitt ved $a_i = X_r^T (\sum_{j \in s} v_j^{-1} x_j x_j^T)^{-1} v_i^{-1} x_i$. Her er $X_r = X - \sum_{i \in s} x_i$, altså summen av x-verdiene til enhetene utenfor utvalget. For en homogen modell vil $X_r^T = N_r$ og $x_j = v_i = 1$ for alle j og i, slik at $a_i = N_r / n$ for alle i. For ratemodellen er $x_i^T = x_i$, og for en enkel lineær regresjonsmodell er $x_i^T = [1 \quad x_i]$. I litteraturen[1] finner vi fire variansestimater for σ_i^2 som er mer robuste enn den som følger av antagelsene om v_i:

- $CV1 = \sum_{i \in s} a_i^2 d_i$, der $d_i = e_i^2$
- $CV2 = \sum_{i \in s} a_i^2 d_i$, der $d_i = e_i^2 / (1 - h_i)$
- $CV2A = \sum_{i \in s} a_i^2 d_i$, der $d_i = e_i^2 \frac{\sum_{j \in s} a_j^2 v_j}{\sum_{j \in s} a_j^2 v_j (1 - h_j)}$
- $CV3 = \sum_{i \in s} a_i^2 d_i$, der $\mathrm{d_i} = \mathrm{e_i^2} / (1 - \mathrm{h_i})^2$

[1] Valliant, R., Dorfman, A.H. og Royall, R.M. (2001). *Finite Population Sampling and Inference: A Prediction approach.* Wiley, New York.

Her er $e_i = y_i - x_i^T\hat{\theta}$ det estimerte residualet. For homogen- og ratemodellen er $\hat{\theta}$ gitt ved henholdsvis (1.1) og (2.1). For en enkel lineær regresjonsmodell er $\hat{\theta} = [\hat{\alpha} \quad \hat{\beta}]^T$, der $\hat{\alpha}$ og $\hat{\beta}$ er gitt ved (3.1) og (3.2). Verdien $h_i = x_i^T(\sum_{j\in s} v_j^{-1}x_j x_j^T)^{-1}v_i^{-1}x_i$ kalles h-verdien til den i-te enheten.

Det andre leddet er det ikke mulig å estimere uten å bruke den antatte variansstrukturen, fordi vi ikke kjenner residualene utenfor utvalget. Det er likevel vanlig å bruke en indirekte estimator gitt ved $\hat{V}(T_r) = \sum_{i\in s} d_i \sum_{i\notin s} v_i / \sum_{i\in s} v_i$. Estimatoren varierer etter hvilken d_i som brukes. Begrunnelsen ligger i å anta at $\sum_{i\notin s} v_i / \sum_{i\in s} v_i \approx \sum_{i\notin s} \sigma_i^2 / \sum_{i\in s} \sigma_i^2$. Denne antagelsen er ikke kritisk så lenge trekkandelen er lav, siden $V(T_r)$ da er mye mindre enn $V(\hat{T}_r)$. For en homogen modell har vi $\hat{V}(T_r) = \sum_{i\in s} d_i N_r / n$ og $\hat{V}(\hat{T}_r) = \sum_{i\in s} d_i N_r^2 / n$, slik at jo større N_r er, jo mindre innflytelse får $V(\hat{T}_r)$ på usikkerheten i prediksjonen. I tilfeller der trekkandelen er stor, må vi bare stole på den antatte variansstrukturen.

A.3 Regresjonsdiagnostikk

I avsnitt 3.5 ser vi på regresjonsdiagnostikk. Struktur lister ut H, R og G, definert på følgende måte:

$$\mathrm{H} = h_{hi} = x_{hi}^T(\sum_{j\in s_h} v_{hi}^{-1}x_{hj}x_{hj}^T)^{-1}v_{hi}^{-1}x_{hj}$$

$$\mathrm{R} = r_{hi} = \frac{\hat{y}_{hj(j)} - y_{hj}}{SD(\hat{y}_{hj(j)} - y_{hj})}$$

$$\mathrm{G} = DFFITS \approx r_{hi}\sqrt{h_{hi}/(1-h_{hi})}$$

H er forklart i A.2, og av formelen ser vi at H er et tall mellom 0 og 1. Dersom $\mathrm{H} > 2p/n$ der p er antall parametere i modellen, antar vi at enheten er avvikende ved at x_i har en uvanlig plassering blant de andre x-ene.

R er den standardiserte differansen mellom estimatet for enhet j når enheten holdes utenfor og den faktiske observasjonen. Hvordan estimatet for enhet j beregnes, avhenger av hvilke modellantagelser som blir gjort. Dersom vi antar at y-ene er identisk normalfordelte vil R være student t-fordelt, med $n_h - 1$ frihetsgrader. Ved store n_h vil t-fordelingen nærme seg normalfordelingen, og derfor brukes kriteriet $|\mathrm{R}| > 2$ for å plukke ut enheter som avviker kraftig fra resten av utvalget ved at residualet er uvanlig stort.

G er en kombinasjon av R og H, og kriteriet $\mathrm{G} > 2\sqrt{p/n}$ brukes for å finne enheter som avviker betydelig fra resten av utvalget, enten på grunn av uvanlig x_i eller stort residual. Dersom H = 0,5 vil G = R. To ting kan forårsake en lav G; H er nær en av grenseverdiene, eller R er lav ved at den observerte verdien av y ligger nær den estimerte verdien.