Statistics Norway
Statistisk sentralbyrå

Li-Chun Zhang

Empirisk imputering

En ny metode for å behandle tilfeldig partielt frafall

Avdeling for samordning og utvikling/Seksjon for statistiske metoder og standarder

Innhold

1 Innledning

Initiativ til dette notatet stammer fra et samarbeid med seksjon for inntekts- og lønnsstatistikk (kontaktperson Bjørg L. Western og Per M. Holt) ang. vekting for inntekts- og formuesundersøkelsen for selskaper 1995 og 1996. Flere fra seksjon for statistiske metoder og standarder (Leiv Solheim, Johan Fosen og Ole Klungsøyr) og Ib Thomsen (avdeling for personstatistikk) har bidratt med nyttige diskusjoner.

Etter en kort innledning om problemstillingen og standard fremgangsmåter, definerer vi empirisk imputering som en metode til behandling av tilfeldig partielt frafall. Dette følges opp av en fyldig beskrivelse av forskjellige testkjøringer, foreløpige resultater, samt sammenligning med andre metoder basert på data fra IFS-95. Det siste avsnittet inneholder oppsummering og en del forslag til videreutvikling.

1.1 Tilfeldig partielt frafall

Partielt frafall er et særdeles utbredt problem i statistisk analyse av data. For eksempel kan en person svare i en utvalgsundersøkelse på noen men ikke alle spørsmål. Et annet vanlig eksempel er når man kobler sammen flere registre, mangler da som regel noen opplysninger hos en del enheter, uansett hva slags populsjon det dreier seg om.

Kommentar 1 *I motsetning betyr enhetsfrafall at det finnes ingen data hos den vedkommende enheten. For eksempel kan en IO nekte å bli intervjuet i det hele tatt.*

Når det ikke finnes spesiell mistanke om systematikken som ligger bak det partielle frafallet, sier man at frafallet er tilfeldig. Spesielt dersom frafallet forekommer uavhengig av verdiene til frafallet.

Kommentar 2 *Modellering av frafallsmekanisme er nødvendig i det tilfellet frafallet ikke er tilfeldig, slik at problemet ikke lenger kan behandles ikke-parametrisk.*

1.2 Registerfrafall i IFS

Datagrunnlaget i dette studiet er hentet inn fra inntekts- og formuesundersøkelsen for selskaper 1995 (IFS-95). Frafallet oppstår her idet man setter sammen opplysninger fra etterskuddsregisteret (**E-reg**), regnskapsregisteret (**BKF**) og det sentrale bedrifts- og foretaksregisteret (**BOF**). Populasjonen i IFS defineres etter E-reg, slik at registertotaler ifølge BKF eller BOF ikke dekker hele populasjonen.

Kommentar 3 *Teknisk sett er det slik at et foretak ikke bidrar til en viss total dersom den relevante opplysningen mangler.*

Til forenkling av problemstillingen skal vi i den følgende teksten konsentere oss kun om registerfrafall i BKF, og se bort fra BOF i det hele tatt. I alt faller c.a. 10% av alle foretakene i E-reg utenfor BKF, mens i selve utvalget IFS-95 er denne andelen noe under 4%.

Kommentar 4 *Følgende diskusjon forusetter noe kjennskap til kalibrering. Leserne som ikke er interessert i å vite detaljer kan gå over til neste avsnittet direkte.*

Den estimeringsmetoden som brukes for IFS kalles kalibrering. Anta generelt p tilleggs variabler $x = (x_1, x_2)$, s.a. x_1 inneholder de q variablene som hentes inn fra E-reg, og x_2 resten $p-q$ variablene fra BKF. Anta i alt N foretak ifølge E-reg. Man kan ordne populasjonen i en $N \times p$-matrise

$$\begin{pmatrix} x_{k1} & x_{k2} \\ - & x_{r2} \end{pmatrix},$$

der $(x_{k1}\ x_{k2})$ står for fortakene i BKF og $(-\ x_{r2})$ for de som er med i E-reg men ikke i BKF. Uten registerfrafall ville man ha fått x_{r1} også. Det viser seg på data fra IFS-95 at kalibrering mht. x_1 har ingen effekt på x_2, dvs. avvikene på den siste registertotalen er like store før og etter kalibrering. Man kan dele estimering for $(x_{k1}\ x_{k2})$ og $(-\ x_{r2})$ hver for seg. Men registerfrafallet der gjør at en del av forklaringsvariablene ble droppet, noe som innfører en skjevhet i metoden. F.eks. i tilfellet med konstante startvekter, er kalibrering med og uten registerfrafallet henholdsvis karakterisert ved

$$x_{r2}(x_{r2}^T x_{r2})^{-1} x_{r2}^T \qquad \text{og} \qquad (x_{r1}\ x_{r2}) \begin{pmatrix} x_{r1}^T x_{r1} & x_{r1}^T x_{r2} \\ x_{r2}^T x_{r1} & x_{r2}^T x_{r2} \end{pmatrix}^{-1} \begin{pmatrix} x_{r1}^T \\ x_{r2}^T \end{pmatrix}.$$

1.3 Noen generelle bemerkninger om imputering

En vanlig metode for behandling av tilfeldig frafall er **hot-deck**: Anta at man har n observasjoner av en viss variabel, og mangler ytterligere m verdier pga. frafall. Da trekker man tilfeldig én observasjon ut ifra de n man har, og setter den som den første manglende verdien, samtidig som den uttrukne verdien legges tilbake blant de n observerte. For å imputere for alle m manglende verdiene, gjentar man prosedyren m ganger.

I tilfellet tilfeldig enhetsfrafall, har hot-deck en del gode egenskaper samtidig som den er enkel i bruk. Den klarer også å bevare den marginale variasjonen i variabelen dersom man har å gjøre med tilfeldig partielt frafall. Men uten å ta hensyn til, eller å være betinget på, de observerte verdiene hos enhetene som inneholder partielt frafall, kommer hot-deck nødvendigvis til å forstyrre kovariasjonstrukturen i multivariate data.

En betinget fremgangsmåte er å lage en **modell** for multivariate data, og imputere så f.eks. betinget forventning for de frafalte komponentene under estimerte parametre. Ved hjelp av gode modeller kan man nå langt med denne metoden. En svakhet ved metoden er det faktum at to enheter med de samme observerte komponentene vil få samme imputering, noe som resulterer i en undervurdering av variasjonen i de imputerte data.

Kommentar 5 *Legg merke til at modellen her ikke er laget for frafallsmekanismen, slik at frafallet fortsatt antas å være tilfeldig.*

En ikke-parametrisk betinget imputeringsmetode heter **nærmeste nabo**: Først bestemmer man en måte å definere den 'nærmeste naboen' til en enhet på. Deretter imputerer man for de frafalte verdiene med de tilsvarende observerte verdiene hos dens nærmeste nabo. I tilfellet den nærmeste nabo er entydig, er også imputeringen entydig. Ellers velger man de imputerte verdiene ved å kjøre hot-deck blant alle sine nærmeste naboer. De fleste anvendelsene av denne metoden går ut på identifisere nærmeste nabo enhetsvis, selv om det ikke alltid er like lett å klargjøre den nærmeste naboen på denne måten gitt multivariate data med partielt frafall.

2 Empirisk imputering

2.1 Definisjon

Anta frafall. For å imputere, eller gjenkonstruere data, må man stille seg det følgende spørsmålet — gitt to forskjellige imputeringer, hvilken ligger nærmest til, eller ligner mest på, observerte data?

Eksempel 1 *Anta data* $(1,1,2,-)$ *med frafall på den siste verdien. Hvilken av* $(1,1,2,1)$ *og* $(1,1,2,2)$ *ligner mest på den observerte* $(1,1,2,-)$*? Er det slik at, med tilfeldig frafall,* $(1,1,2,1,1,2)$ *ligger nærmest til* $(1,1,2,-,-,-)$ *blant alle imputeringer?*

Anta $y = (y_1, ..., y_n)$, muligens multivariate. Den **empiriske tetthetsfunksjonen** tilordner sannsynnligheten $1/n$ til hver av disse n observasjonene, s.a. den **empiriske fordelingsfunksjonen** basert på y er

$$\hat{F}_Y(z) = \hat{P}[Y \leq z] = \frac{\#\{y_i; y_i \leq z\}}{n}, \tag{1}$$

der telleren er antall observasjoner mindre eller lik z — for multivariate y betyr $y_i \leq z$ at alle komponentene til y_i er mindre eller lik de tilsvarende komponentene til z. All informasjon som ligger i data er nå oppsummert i den empiriske fordelingsfunksjonen.

Siden analysen sikter på å trekke informasjon ut av data, og den empiriske fordelingsfunksjonen summerer opp all informasjon i data, foreslår vi å måle avstanden mellom to sett av data mht. deres empiriske fordelingsfunksjoner. Betegn det observerte datasettet med y og det imputerte med x; betegn den empiriske fordelingsfunksjonen basert på y med $\hat{F}_Y$, og den på x med $\hat{F}_X$. Vi definerer δ-**verdien** på imputering x, i forholdet til oberverte y, som

$$\delta(x) = \int |\hat{F}_X(z) - \hat{F}_Y(z)| \, dz, \tag{2}$$

s.a. den svarer til arealet mellom $\hat{F}_X$ og $\hat{F}_Y$ dersom x og y er univariate. Vi skal bruker denne δ-verdien som avstanden mellom to sett av data. Da er en **empirisk imputering** per definisjon s.a. ingen andre imputeringer har mindre δ-verdi i forholdet til oberverte data.

Eksempel 2 *Anta* $y = (1, 1, 2)$ *og* $x = (1, 1, 2, 1)$ *s.a.*

$$\hat{F}_Y = \begin{cases} 0 & z < 1 \\ 2/3 & 1 \leq z < 2 \\ 1 & z \geq 2 \end{cases} \qquad \textit{og} \qquad \hat{F}_X = \begin{cases} 0 & z < 1 \\ 3/4 & 1 \leq z < 2 \\ 1 & z \geq 2 \end{cases}$$

og $\delta(x) = |3/4 - 2/3| \cdot 1 = 1/12$. *Leserne kan sjekke selv at* $\delta(x) = 1/6$ *med samme* y *og* $x = (1, 1, 2, 2)$, *og* $\delta(x) = 0$ *hvis* $x = (1, 1, 2, 1, 1, 2)$.

Eksempel 3 *Anta* $y = c(12.8, 13.1, 13.2, 13.4, 15)$ *og at det er kun én manglende verdi man skal imputere. La* $x = (x_1, ..., x_6)$ *der* $(x_1, ..., x_5) = y$, *og*

$$\delta(x) = \begin{cases} |\frac{2}{6} - \frac{1}{5}| \cdot 0.3 + |\frac{3}{6} - \frac{2}{5}| \cdot 0.1 + |\frac{4}{6} - \frac{3}{5}| \cdot 0.2 + |\frac{5}{6} - \frac{4}{5}| \cdot 1.6 = 0.117 & x_6 = 12.8, \\ |\frac{1}{6} - \frac{1}{5}| \cdot 0.3 + |\frac{3}{6} - \frac{2}{5}| \cdot 0.1 + |\frac{4}{6} - \frac{3}{5}| \cdot 0.2 + |\frac{5}{6} - \frac{4}{5}| \cdot 1.6 = 0.087 & x_6 = 13.1, \\ |\frac{1}{6} - \frac{1}{5}| \cdot 0.3 + |\frac{2}{6} - \frac{2}{5}| \cdot 0.1 + |\frac{4}{6} - \frac{3}{5}| \cdot 0.2 + |\frac{5}{6} - \frac{4}{5}| \cdot 1.6 = 0.083 & x_6 = 13.2, \\ |\frac{1}{6} - \frac{1}{5}| \cdot 0.3 + |\frac{2}{6} - \frac{2}{5}| \cdot 0.1 + |\frac{3}{6} - \frac{3}{5}| \cdot 0.2 + |\frac{5}{6} - \frac{4}{5}| \cdot 1.6 = 0.09 & x_6 = 13.4, \\ |\frac{1}{6} - \frac{1}{5}| \cdot 0.3 + |\frac{2}{6} - \frac{2}{5}| \cdot 0.1 + |\frac{3}{6} - \frac{3}{5}| \cdot 0.2 + |\frac{4}{6} - \frac{4}{5}| \cdot 1.6 = 0.25 & x_6 = 15, \end{cases}$$

s.a. $x_6 = 13.2$ *gir den empiriske imputeringen.*

Kommentar 6 *Som man kanskje allerede har lagt merke til, det er noe kunstig med imputering for tilfeldig enhetsfrafall. På en måte kan man tenke alle de uobserverte enhetene som tilfeldig enhetsfrafall, s.a. problemet er enten det samme som prediksjon for endelig antall enheter, eller estimering for uendelig super-populasjon.*

2.2 Trinnvis empirisk imputering — den grådige algoritmen

Det er enkelt å beregne (2) i univariate tilfeller. Integrering over multidimensjonale rom kan gjøres via betingete univariate fordelinger. Anta bivariat fordeling $F(x_1, x_2)$, der x_1 alltid observeres. (Legg merke til at fot-indeks her står for komponenter istedenfor enheter.) Mao. er den marginale empiriske fordelingen til X_1 bestemt. Dette medfølger at δ-verdien er minimert dersom den er minimert betinget på alle verdier til X_1. Disse betingete δ-verdiene er alle univariate.

Generelt trenger heller ikke x_1 alltid å være observert. Man kan f.eks. begynne med en hot-deck imputering for x_1 som vanlig, og deretter kjøre empirisk imputering for x_2 betinget på imputert x_1, og iterere.

Kommentar 7 *En slik algoritme ligner på såkalt Gibbs-sampler.*

Mer komplisert er å danne betinget empirisk fordeling. Anta bivariate $(y_{j,1}, ..., y_{j,n}, y_{j,n+1})$, $j = 1, 2$, der $y_{1,i}$ er alle observert for $i = 1, ..., n+1$, mens $y_{2,n+1}$ er tilfeldig frafall. Dersom $y_{1,n+1}$ er lik en eller flere blant $(y_{1,1}, ..., y_{1,n})$, er det klart at disse enhetene skal brukes for den empiriske fordelingen til Y_2 betinget på $Y_1 = y_{1,n+1}$. Generelt, også når $\nexists 1 \leq i \leq n$ s.a. $y_{1,i} = y_{1,n+1}$, kan man bruke de enhetene som har y_1-verdier i nærheten av $y_{1,n+1}$. For eksempel kan man inkludere alle enheter med $(1-\alpha)y_{1,n+1} \leq Y_1 \leq (1+\alpha)y_{1,n+1}$, eller $y_{1,n+1} - \epsilon \leq Y_1 \leq y_{1,n+1} + \epsilon$, der $\alpha \geq 0$. Mao. lager man et **vindu** for $y_{1,n+1}$ og baserer den betingete empiriske fordelingen på alle enhetene som faller inn i dette vinduet.

Kommentar 8 *Et slikt vindu er vanlig i ikke-parametrisk tetthetsestimering, der det heter "band-width". Brede vinduer glatter ut hakk i de observerte verdiene, mens smale vinduer gjør betingete fordelinger mer følsomme overfor lokale variasjoner i data.*

Kommentar 9 *Selv om alle enheter med partielt frafall kan få individuelle vinduer, blir oppgavebyr-den fort umenneskelig dersom de ikke lages etter noen enkle regler. Dette betyr av og til at data må transformeres til en mer passende skala før imputering settes i gang.*

Kommentar 10 *Siden kun verdiene som faller inn i et eller annet vindu er med på å danne de empiriske betingete fordelingene, bruker metoden som regel ikke alle de observerte verdiene. Dette ligner på nærmeste nabo der man kun benytter verdiene til dem som er nærmeste naboer hos en eller annen enhet med partielt frafall.*

Anta til slutt bivariate $(y_{1,i}, y_{2,i})$, $i = 1, .., n+2$, med partielt frafall på $(y_{2,n+1}, y_{2,n+2})$, og at $y_{1,n+1} = y_{1,n+2}$, s.a. de to refererer til den samme betingete fordelingen. Istedenfor å imputere to verdier samtidig, kan man ta en av dem om gangen, så lenge den andre imputerte verdien er beregnet med hensyn til endring på fordelingen pga. av den første imputerte verdien. Man kan kalle dette en **grådig algoritme**, siden den optimaliserer på hvert trinn, men uten nødvendigvis å treffe det helhetlige optimum.

Kommentar 11 *Den grådige algoritmen er en enkel metode, som bryter letingen etter empirisk imputering for multidimensjonalt partielt frafall ned til enkle komponentvise δ-verdi beregninger som i eksemplene tidligere.*

Eksempel 4 *Anta igjen* $y = c(12.8, 13.1, 13.2, 13.4, 15)$ *og at man skal imputere to manglende verdier. Etter å ha sjekket δ-verdi til alle 15 mulige kombinasjon av de to imputerte verdiene, kunne man slå fast at* $(13.1, 13.4)$ *gir den minste δ-verdien på 0.117. Samtidig gir den grådige algoritmen*

13.2 på det første trinnet (vist tidligere), og 13.4 på det andre betinget på 13.2 som den første imputerte verdien. Sammen gir $(13.2, 13.4)$ *en* δ*-verdi på 0.126 som, i likhet med* $(12.8, 13.4)$*, gir den nest minste* δ*-verdien blant alle de mulige imputerte verdiene.*

Videre gir den grådige algoritmen henholdsvis $(13.2, 13.4, 15)$ *og* $(13.2, 13.4, 15, 12.8)$ *når antall manglende verdier er 3 og 4. De faktiske empiriske imputeringene i de to tilfellene er derimot* $(12.8, 13.2, 15)$ *med en* δ*-verdi på 0.1025, og* $(12.8, 13.1, 13.4, 15)$ *med en* δ*-verdi på 0.056. Den grådige algoritmen treffer den empiriske imputeringen når antall manglende verdier er 5, nemlig* $(13.2, 13.4, .15, 12.8, 13.1)$*, med* δ*-verdi 0.*

Vi noterer f.eks. at når antall manglende verdier er 4, finnes det i alt 70 mulige kombinasjoner av imputerte verdier. Mens den grådige algoritmen trenger bare å beregne 20 $(= 5 \cdot 4)$ δ*-verdier.*

2.3 Tilleggs bemerkninger

Empirisk imputering fremstår som en slags kanonisert nærmeste nabo imputering der den siste kan virke *ad-hoc*. Identifikasjon av den nærmeste naboen kan nå gis en mer presis formulering gjennom betinget fordeling, med sikte på å komme nærmest mulig til den observerte multivariate empiriske fordelingen.

Som en optimaliseringsprosedyre er empirisk imputering i utgangspunktet entydig. Men pga. endeligheten i bestemte datasett, hender det av og til at flere imputeringer har den samme minste δ-verdien. I slike tilfeller kan man bruke hot-deck blant disse.

Eksempel 5 *Anta* $y = (1, 2, -)$*. Da er man nødt til å bruke hot-deck.*

Det er viktig her at man ikke forveksler entydigheten i empirisk imputering med underslått variasjon i data. Ved å gjøre den imputerte empiriske fordelingen tilnærmet det samme som den observerte empiriske fordelingen, har metoden sikret *all informasjon i data*, inkl. variasjon i data.

Man skal heller ikke forveksle variasjonen i data med usikkerheten til metoden. Det første hører til hva data sier, mens den siste handler om hvor sikkert data sier det data sier. Man vil undervurdere usikkerheten i metoden dersom man baserer vurderingen på imputerte data uten å ta hensyn til usikkerheten i imputering, rett og slett fordi det imputerte datasettet er større enn det observerte. Men dette betyr ikke nødvendigvis at variasjonen er mindre i imputerte data enn i observerte data.

Eksempel 6 *Anta* $y = (1, 1, 2, -, -, -)$ *og den empiriske imputeringen* $x = (1, 1, 2, 1, 1, 2)$*. De to gir den samme empiriske fordelingen. Dette innebær bl.a. at de har den samme gjennomsnittet, nemlig* $\bar{y} = \bar{x} = 4/3$*, og den samme variasjonen i data, nemlig* $\sum_{i=1}^{3}(y_i - \bar{y})^2/3 = \sum_{j=1}^{6}(x_j - \bar{x})^2/6 = 2/9$*. På den andre siden er det slik at* $\bar{Y}$ *som estimatoren for forventningen har omt. dobbelt så stor usikkerhet som* $\bar{X}$*. Dette vet man egentlig ut ifra at datasettet* X *er dobbelt så stort som* Y*, uten at man trenger å kjenne til* (x, y)*, dvs. verdiene i datasettene.*

3 Empirisk imputering for data fra IFS-95 (I)

3.1 Transformajson av data

Variablene vi skal behandle i denne teksten deles i to grupper. De første fem av dem hentes inn fra BKF, nemlig (1) **BKF-DRI**: sum driftsinntekter, (2) **BKF-DRES**: driftsresultat, (3) **BKF-ARSR**: netto årsresultat, (4) **BKF-EIEN**: sum eiendeler, og (5) **BKF-AKAP**: aksjekapital. Partielt frafall skjer på disse fem variablene dersom et foretak i E-reg ikke befinnes i BKF. De siste to hentes inn fra E-reg, nemlig (6) **ASKAP-BR**: aksjekapital ifølge E-reg, og (7) **ALM**: alminnelige inntekter. Disse

to skal være kjent i hele populasjonen. Videre er BKF-DRI, BKF-EIEN, BKF-AKAP og ASKAP-BR ikke-negative, mens BKF-DRES, BKF-ARSR og ALM kan være både positive og negative.

Kommentar 12 *BKF-DRI er et slags bruttoresultat, og forskjellen mellom den og BKF-DRES, eller BKF-ARSR, regnes som forskjellige kostnader. Mens skillet mellom BKF-AKAP og ASKAP-BR ligger i at de to er hentet inn på forskjellige tidspunkter.*

Som regel forstyrres slike data av ekstreme verdier i halene til fordelingen. I tilfellet data, si y, er strengt positive, kan man ta en log-transformasjon direkte, nemlig $\log(y+1)$, som ikke bare bevarer fortegnet på data, men også behandler null på en elegant måte siden $\log(0+1)=0$. Generelt kan man vurdere

$$z = \log(a \cdot y - b \cdot \min(y) + c)$$

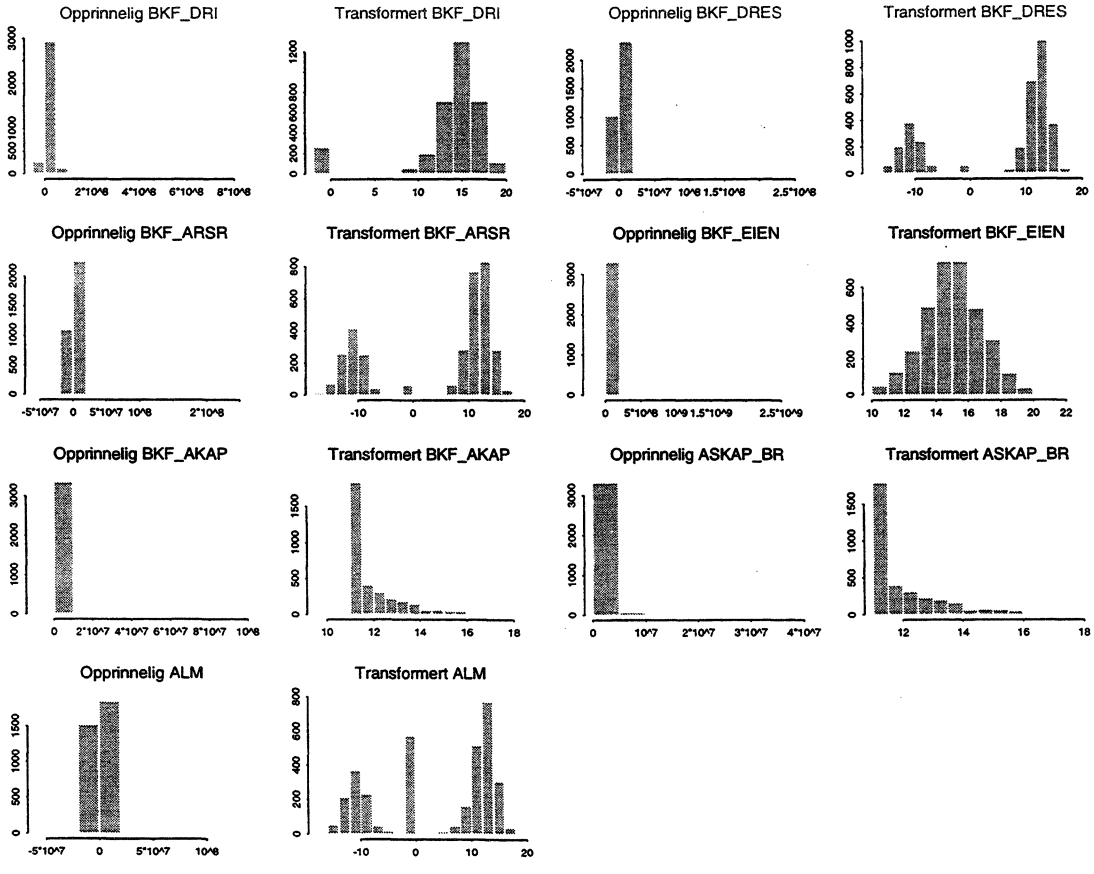
for fastsatte (a, b, c), eller, for fastsatte (a, b),

$$z = \frac{y}{|y|}\log(a \cdot |y| + b) \quad \text{hvis } y \neq 0 \qquad \text{og} \qquad z = 0 \quad \text{hvis } y = 0. \tag{3}$$

Kommentar 13 *Den første transformasjonen har av og til en tendens til å 'klemme' de fleste data (i midten) inn i et veldig lite intervall. Mens den andre gjør at data ville få tre topper — en positv, en negativ og en på null. Uansett er målet her bl.a. å minske effekten av ekstreme data på analysen.*

Uten spesifikasjon ellers, skal vi fra nå av benytte transformasjon (3) med konstant

	BKF-DRI	BKF-DRES	BKF-ARSR	BKF-EIEN	BKF-AKAP	ASKAP-BR	ALM
a	1	1	1	2	1	1	1
b	1	1	1	5×10^4	2×10^4	2×10^4	1



I alt består data av 3545 foretak, deriblant 162 frafall i BKF. Histogram ovenfor er basert på de 3383 fortak uten registerfrafall, som bl.a. viser at disse transformasjonene demper de ekstreme verdiene slik at data har fått en bedre marginal spredning. Samtidig er det i tabellen nedenfor listet ut parvise korrelasjoner før og etter transformasjonene.

	Korrelasjon før transformasjon					
	BKF-DRES	BKF-ARSR	BKF-EIEN	BKF-AKAP	ASKAP-BR	ALM
BKF-DRI	0.48	0.41	0.51	0.23	0.36	0.33
BKF-DRES		0.86	0.81	0.24	0.45	0.47
BKF-ARSR			0.70	0.19	0.36	0.64
BKF-EIEN				0.25	0.38	0.42
BKF-AKAP					0.59	0.10
ASKAP-BR						0.32
	Korrelasjon etter transformasjon					
	BKF-DRES	BKF-ARSR	BKF-EIEN	BKF-AKAP	ASKAP-BR	ALM
BKF-DRI	0.43	0.27	0.52	0.20	0.19	0.24
BKF-DRES		0.76	0.33	0.06	0.06	0.62
BKF-ARSR			0.31	0.10	0.11	0.73
BKF-EIEN				0.59	0.58	0.33
BKF-AKAP					0.93	0.08
ASKAP-BR						0.09

Hvis man ser langs kolonner for ASKAP-BR og ALM, dvs. de to variablene som er uten fare for registerfrafall, kan det virke som ASKAP-BR er best korrelert med BKF-AKAP og BKF-EIEN etter transformasjonen, og ALM med BKF-DRES og BKF-ARSR. Mao. en seperasjon mellom de første tre variablene og de siste tre, noe som kanskje også er naturlig fra definisjonen av disse variablene.

Kommentar 14 *Teknisk sett gir en slik gruppering blant variablene muligheten til å redusere en høy-dimensjonal simultan fordeling til flere lavere dimensjonale fordelinger.*

3.2 Tilfeldig partielt frafall?

Det er som regel lønnsomt å se litt nærmere på data før man setter igang analysen. Det er svært viktig i denne sammenhengen å avklare om frafallet i BKF er tilfeldig. Først sjekker vi om registerfrafallet er næringbestemt. Resultatet er som følgende:

	Næringsfordeling blant registerfrafall								
Næring	Ukjent	I	II	III	IV	V	VI	VII	IX
Frafall	34	14	9	45	14	8	33	4	1
I alt	231	347	359	1057	244	173	922	148	64
Andel (%)	15	4	3	4	6	5	4	3	2

Forståelig nok er frafallet litt høyere blant foretak uten nærings opplysning. Ellers er frafallet ganske jevnt fordelt over alle næringer — næring VIII er per definisjon tom.

Med "kjerne-data" refererer vi til de 3383 foretak uten frafall og "brutto-data" til alle 3545 foretak. Beregning viser at $X = (\text{ASKAP-BR}, \text{ALM})$ har gjennomsnittet

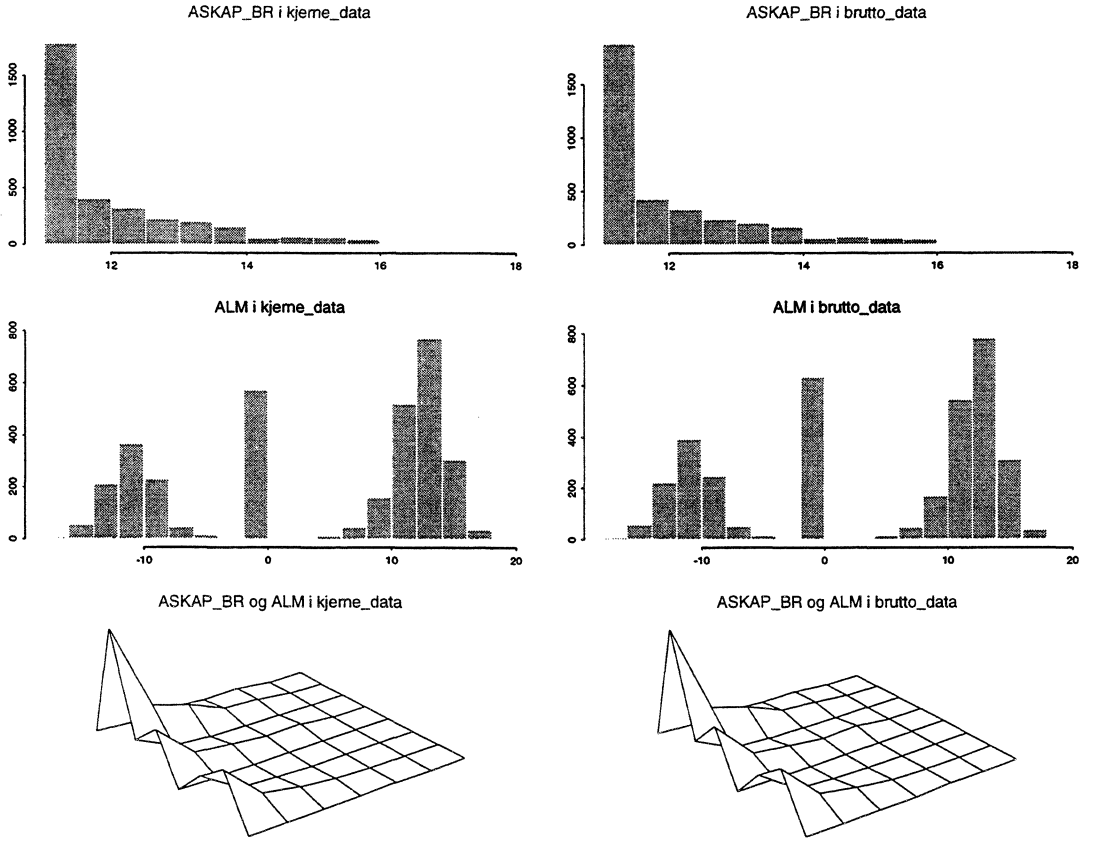
$$\bar{X}_{kjerne} = (11.986,\ 3.696) \qquad \bar{X}_{brutto} = (11.979,\ 3.515),$$

og kovariansen

$$Cov_{kjerne}(X) = \begin{pmatrix} 1.40 & 1.11 \\ 1.11 & 106.53 \end{pmatrix} \qquad Cov_{brutto}(X) = \begin{pmatrix} 1.39 & 1.06 \\ 1.06 & 106.02 \end{pmatrix}.$$

Kommentar 15 *Variansen til transformert ALM er ikke nødvendigvis et godt mål på den marginale variasjonen i data pga. flere topper i fordelingen, heller ikke på den opprinnelige skalaen pga. alle ekstreme verdier.*

Samtidig viser histogrammene under fordelingene til de transformerte ASKAP-BR og ALM, både de marginale og den simultane fordelingen.



3.3 Simulerte data

Undersøkelse av data ovenfor tyder på at det ikke er helt urealistisk å anta at frafallet i BKF er tilfeldig. For å studere metoden, skal vi imidlertid lage et tilfeldig datasett. Mao. har vi trukket ut i alt **180** enheter fra de **3383** foretak som er uten registerfrafall, og merket disse som frafallet, s.a. simulerte kjerne data består av **3203** foretak. Frafallsraten målt i antall enheter er derfor $180/3383 = 5.3\%$.

Kommentar 16 *Ikke bare er dette simulerte datasettet tilrettelagt med antagelsen om tilfeldig frafall, det inneholder også fasitten til verdiene i frafallet.*

Uten spesifikasjon ellers, skal vi bruke disse simulerte data i det følgende studiet.

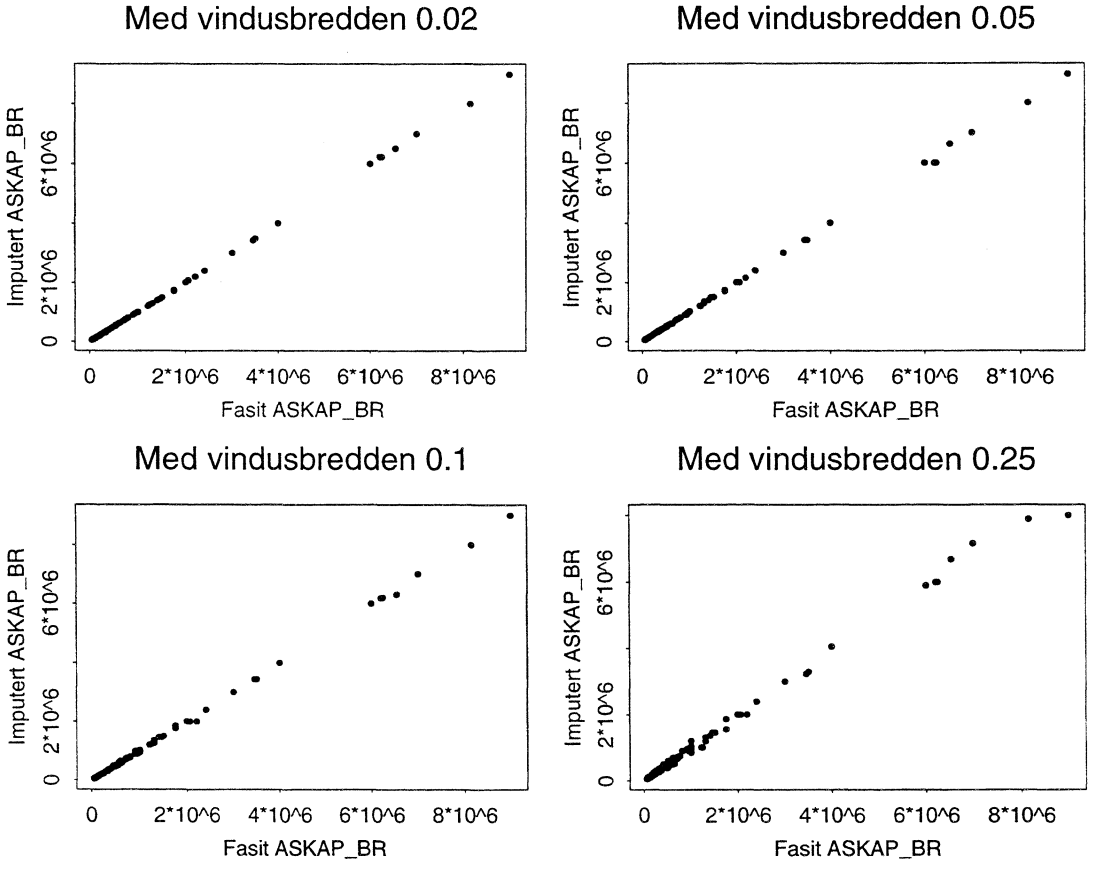
3.4 Valg av vindusbredden

Vi skal forsøke å utnytte grupperingen blant variablene nevnt tidligere med å imputere for BKF-EIEN og BKF-AKAP betinget på ASKAP-BR, og BKF-DRES og BKF-ARSR på ALM. På mange måter dreier det vanskeligste valget her seg om vindusbredden, noe som avgjør hvilke betingete fordelinger man skal basere empirisk imputering på. En opplagt fremgangsmåte å prøve ut valget på selve ASKAP-BR og ALM, og late som om man ikke har fasitten på de frafalte BKF-verdiene.

Betrakt først ASKAP-BR. Frafallsraten målt i X = ASKAP-BR er $(\sum_{j\in s_f} x_j)/(\sum_{i\in s} x_i) = 8.6\%$ på opprinnelige skala og 5.7% på transformerte data, der s betegner utvalget og s_f dets frafallsdel. For $j \in s_f$ med ASKAP-BR-verdi x_j, lager vi et α-vindu som

$$\text{Vindu}_\alpha(j) = \{i \in s_s; \quad x_j - \alpha \leq x_i \leq x_j + \alpha, \quad \alpha > 0\}, \tag{4}$$

der s_s betegner den komplette delen av utvalget. Legg merke til et (absolutt) α-vindu på den log-skala er omt. lik et (relativt) $100 \cdot \alpha\%$-vindu på den opprinnelige skala for små α. For å finne empirisk imputering skal vi, for hver $j \in s_f$, minimere δ-verdi over Vindu(j) mht. den empiriske fordelingen basert på $\{x_i; i \in \text{Vindu}_\alpha(j)\}$. Vi har plottet de imputerte verdiene mot fasitten for alle 180 foretakene ved forskjellige α.



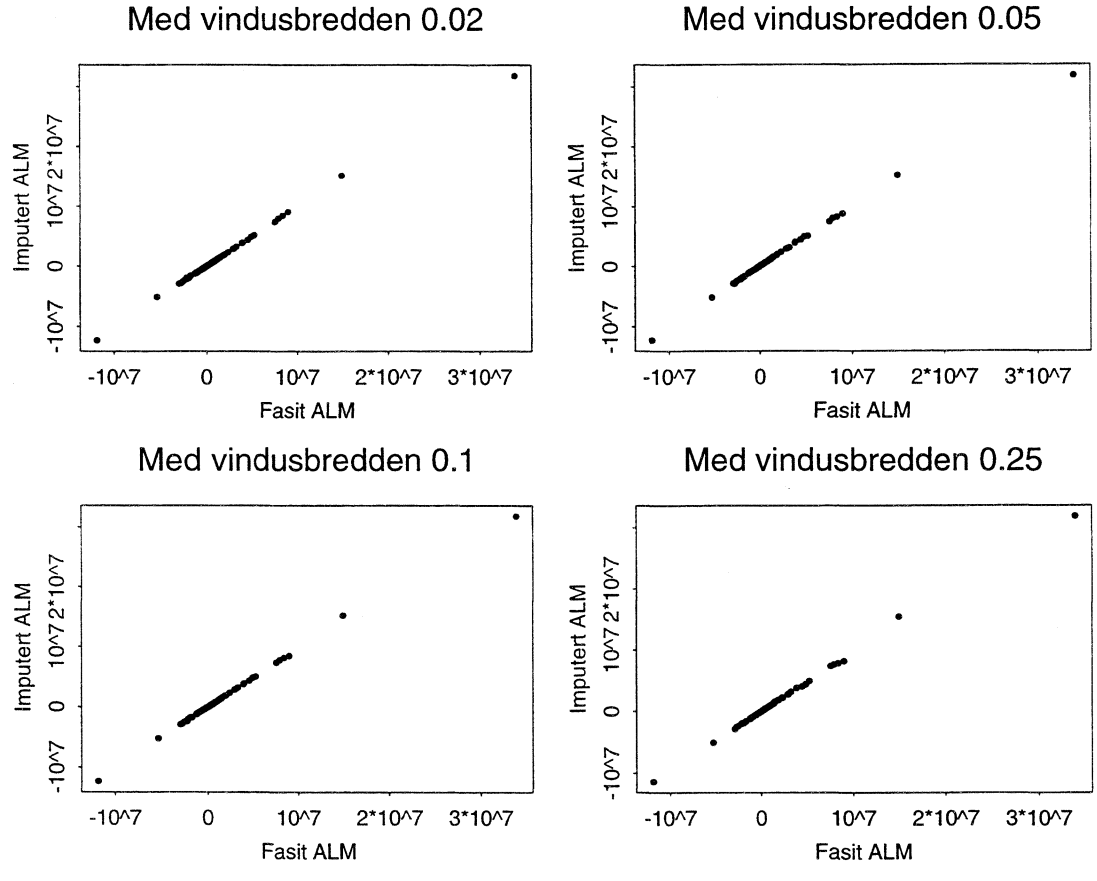
Kommentar 17 *Det var ett tilfelle med tomt vindu når $\alpha = 0.02$. For akkurat denne enheten har vi satt vindusbredden til 0.20. En forholdsvis stor vindusbredde forvandler metoden til å ligne mer på komponentvis hot-deck i slik tilfelle.*

Empirisk imputering kommer her til å ligne mer og mer på nærmeste nabo ettersom vindusbredden blir smalere og smalere, gitt at man kan finne alle de frafalte verdiene blant de observerte. Dette

betyr at en slik øvelse, nemlig å bruke ASKAP-BR til å imputere for ASKAP-BR, er mer informativ om den øvre grensen til vindusbredden enn den nedre. Mens $\alpha = 0.25$ kan begynne å virke for stor, er valget mellom $\alpha = 0.05$ og 0.10 ikke så lett å avgjøre. Alt i alt synes resultatene å være rimelig robuste overfor valg av vindusbredde. Følgende tabell viser en del detaljer i tillegg.

	Enheter	Transformert			Opprinnelig		
α	brukt (%)	Sum-fasit	Sum-imput	Andel (%)	Sum-fasit	Sum-imput	Andel (%)
0.02	92.7	2302	2302	100.0	136119775	135852737	99.8
0.05	97.1	2302	2301	100.0	136119775	135321157	99.4
0.10	98.7	2302	2301	100.0	136119775	135440559	99.5
0.25	99.9	2302	2299	99.9	136119775	132848570	97.6

Samme øvelsen for ALM gir temmelig like resultater. (Vi noterer at frafallsraten målt i $X =$ ALM er 6.9% på den opprinnelige skala og 5.0% på transformerte data.)



Kommentar 18 *Det var henholdsvis seks, to og ett tilfeller med tomme vinduer når $\alpha = 0.02$, 0.05 og 0.10. I alle disse tilfellene satte vi vindusbredden til 0.20 istedet.*

	Enheter	Transformert			Opprinnelig		
α	brukt (%)	Sum-fasit	Sum-imput	Andel (%)	Sum-fasit	Sum-imput	Andel (%)
0.02	44.1	628	628	100.0	128930600	127417900	98.8
0.05	65.3	628	628	99.9	128930600	125945500	97.7
0.10	83.7	628	628	99.9	128930600	125190100	97.1
0.25	93.3	628	628	99.9	128930600	124075300	96.2

4 Empirisk imputering for data fra IFS-95 (II)

4.1 Opplegget og alternative metoder

Vi skal beregne empirisk imputering for alle fem BKF-variablene på følgende måte: (1) imputer henholdsvis for BKF-EIEN og BKF-AKAP mht. simultan empirisk fordeling (BKF-EIEN, ASKAP-BR) og (BKF-AKAP, ASKAP-BR), (2) imputer henholdsvis for BKF-DRES og BKF-ARSR mht. simultan empirisk fordeling (BKF-DRES, ALM) og (BKF-ARSR, ALM), (3) imputer for BKF-DRI mht. simultan empirisk fordeling (BKF-DRI, ASKAP-BR). Spesielt refererer vi til empirisk imputering med vindusbredden $\alpha = 0.05$ som **Emp-I** og $\alpha = 0.10$ som **Emp-II**.

Kommentar 19 *Dette er på mange måter det enkleste opplegget. Formålet er først og fremst å studere metoden framfor å lage et produksjonsopplegg. F.eks. bruker vi ASKAP-BR for BKF-DRI mest pga. at begge to er ikke-negative. Samtidig er det også interessant å se hvordan metoden fungerer for slike svakt korrelerte variabler — korrelasjonskoeffisienten mellom transformerte BKF-DRI og ASKAP-BR er ikke mer enn 0.19.*

Til sammenligning skal vi også se på betinget forventning imputering under enkel lineær regresjon. Ta f.eks. imputering for BKF-EIEN mht. simultan fordeling (Y, X) = (BKF-EIEN, ASKAP-BR). Regresjon med normalfordelt restledd er da

$$y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \quad \text{for } i = 1, ..., 3383. \tag{5}$$

Spesielt kan (y, x) være på den opprinnelige eller transformerte skala. Imputering finner man ved først å tilpasse modellen basert på 3203 foretak uten registerfrafall, deretter å beregne betinget forventning (gitt x) for de 180 foretak under de estimerte parametrene. I tilfellet regresjon på log-skala, må man videre transformere de imputerte verdiene tilbake til den opprinnelige skala. Den samme prosedyren brukes for alle de andre BKF-variablene. Til referanse skal vi kalle regresjon på transformerte data **Reg-I** og regresjon på den opprinnelige skalaen **Reg-II**.

Kommentar 20 *I motsetning til empirisk imputering bruker betinget forventning imputering alle enheter uten frafall. Den erstatter alle de betingete fordelingene med én normalfordeling, og all avhengighet mellom respons og de uavhengige variablene blir sammenfattet i den lineære prediktoren.*

I tilegg skal vi ta med to *ad hoc* metoder. Den ene baserer seg på de 3203 foretak uten frafall og ser totalt bort fra de resten 180 foretak, som vi kaller **Met-I**. Mens den andre imputerer ganske enkelt verdien null på den opprinnelige skalaen i alle tilfeller, nemlig **Met-0**.

Kommentar 21 *Teknisk sett svarer dagens behandling av register-verdier til Met-0. Som f.eks. når man kjører ut register-totaler for BKF-variablene, vil de frafalte verdiene bli talt opp som null.*

4.2 Resultater (I)

Vi oppsummerer i tabellen de marginale resultatene, dvs. imputerte mot fasitten, beregnet på alle 3383 foretak, med unntak på Met-I som er basert på de 3203 foretak uten frafall, og noterer det følgende: (1) Emp-I og Emp-II skiller seg veldig lite fra hverandre på transformerte data. På den opprinnelige skala er imidlertidig $\alpha = 0.1$ best i tilfellet BKF-DRES, og omvendt i tilfellet BKF-ARSR. Overalt synes empirisk imputering er være rimelig robust overfor valget av vindusbredden. (2) Sammenlignet med Reg-I, som er beregnet på log-skala som empirisk imputering, er det nesten

ingen forskjell mellom de to metodene. Men når man transformerer de imputerte verdiene tilbake til den opprinnlige skalaen, er empirisk imputering klart best. (3) Reg-II, som er beregnet direkte på den opprinnelige skala, viser den samme sårbarheten overfor valget av skala. Av og til er den ustabil — trolig pga. ekstreme verdier, som i tilfellet BKF-ARSR der Reg-II faktisk er dårligere enn Met-0. I motsetning er empirisk imputering klart bedre enn Met-0 hele veien. (4) Met-I ser helt bort fra frafallsdelen. Marginalt sett gir den derfor resultater som er omtrent som hva man kan forvente av komponentvis hot-deck. Stort sett har empirisk imputering klart seg innenfor disse feilmarginene som er satt av tilfeldighet, selv i tilfellet BKF-DRI der korrelasjonen er svak.

	Transformerte data						
Gj.-snitt	Emp-I	Emp-II	Reg-I	Reg-II	Fasiten	Met-I	Met-0
BKF-DRI	13.626	13.631	13.586	13.701	13.632	13.555	12.834
Relat. Diff.	0.000	0.000	-0.002	0.005	0	-0.005	-0.058
BKF-DRES	5.425	5.422	5.418	5.711	5.466	5.425	5.137
Relat. Diff.	-0.007	-0.007	-0.008	0.045	0	-0.006	-0.059
BKF-ARSR	4.538	4.543	4.536	4.656	4.589	4.545	4.303
Relat. Diff.	-0.01	-0.009	-0.01	0.015	0	-0.009	-0.061
BKF-EIEN	15.04	15.04	15.034	15.08	15.033	14.994	14.772
Relat. Diff.	0.000	0.000	0.000	0.003	0	-0.002	-0.016
BKF-AKAP	11.958	11.959	11.957	11.974	11.955	11.914	11.807
Relat. Diff.	0.000	0.000	0.000	0.002	0	-0.002	-0.011
	Opprinnelige skala						
Gj.-snitt	Emp-I	Emp-II	Reg-I	Reg-II	Fasiten	Met-I	Met-0
BKF-DRI	11719260	11756266	11293678	11991261	12449732	11818790	11189945
Relat. Diff.	-0.058	-0.055	-0.092	-0.036	0	-0.05	-0.1
BKF-DRES	633058	638547	607063	641301	644319	635790	601962
Relat. Diff.	-0.016	-0.008	-0.057	-0.004	0	-0.012	-0.065
BKF-ARSR	466903	474686	465731	490901	469220	483328	457611
Ralat. Diff.	-0.004	0.012	-0.006	0.046	0	0.03	-0.024
BKF-EIEN	9269851	9258049	9175868	9605065	9480703	9339959	8843006
Ralat. Diff.	-0.021	-0.022	-0.031	0.013	0	-0.014	-0.066
BKF-AKAP	570244	570655	565431	579721	572254	560971	531123
Ralat. Diff.	-0.003	-0.002	-0.011	0.013	0	-0.019	-0.071

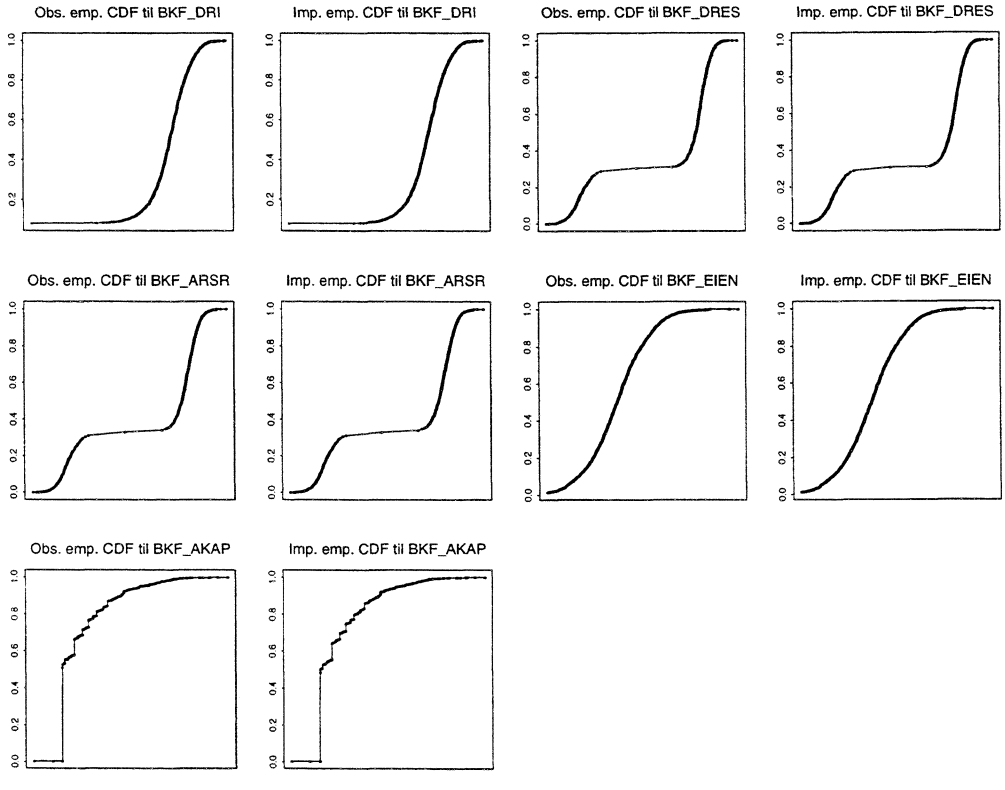
Resultater mht. variasjon i data oppsumerer vi i tabellen som følger. I tillegg noterer vi at: (1) varians på den opprinnelige skala er et dårlig mål for variasjon i data — den klarte ikke en gang å avsløre Met-0 som en dårlig metode i denne sammenhengen. (2) Empirisk imputering har vist robustheten overfor valget av vindusbredden. (3) Empirisk imputering er bedre enn Reg-I i denne sammenhengen. (4) Empirisk imputering er også stort sett bedre enn Met-I.

Kommentar 22 *Det at empirisk imputering stort sett har gitt bedre resultater enn Met-I betyr at det er nyttig å satse på imputering selv når frafallet er tilfeldig, siden det tross alt ligger informasjon i de observerte verdiene hos enhetene med partielt frafall.*

Til slutt skal vi ta med oss alle de marginale empiriske fordelingene før (3203 foretak) og etter (3383 foretak) imputering på transformerte data, med vindusbredden $\alpha = 0.05$.

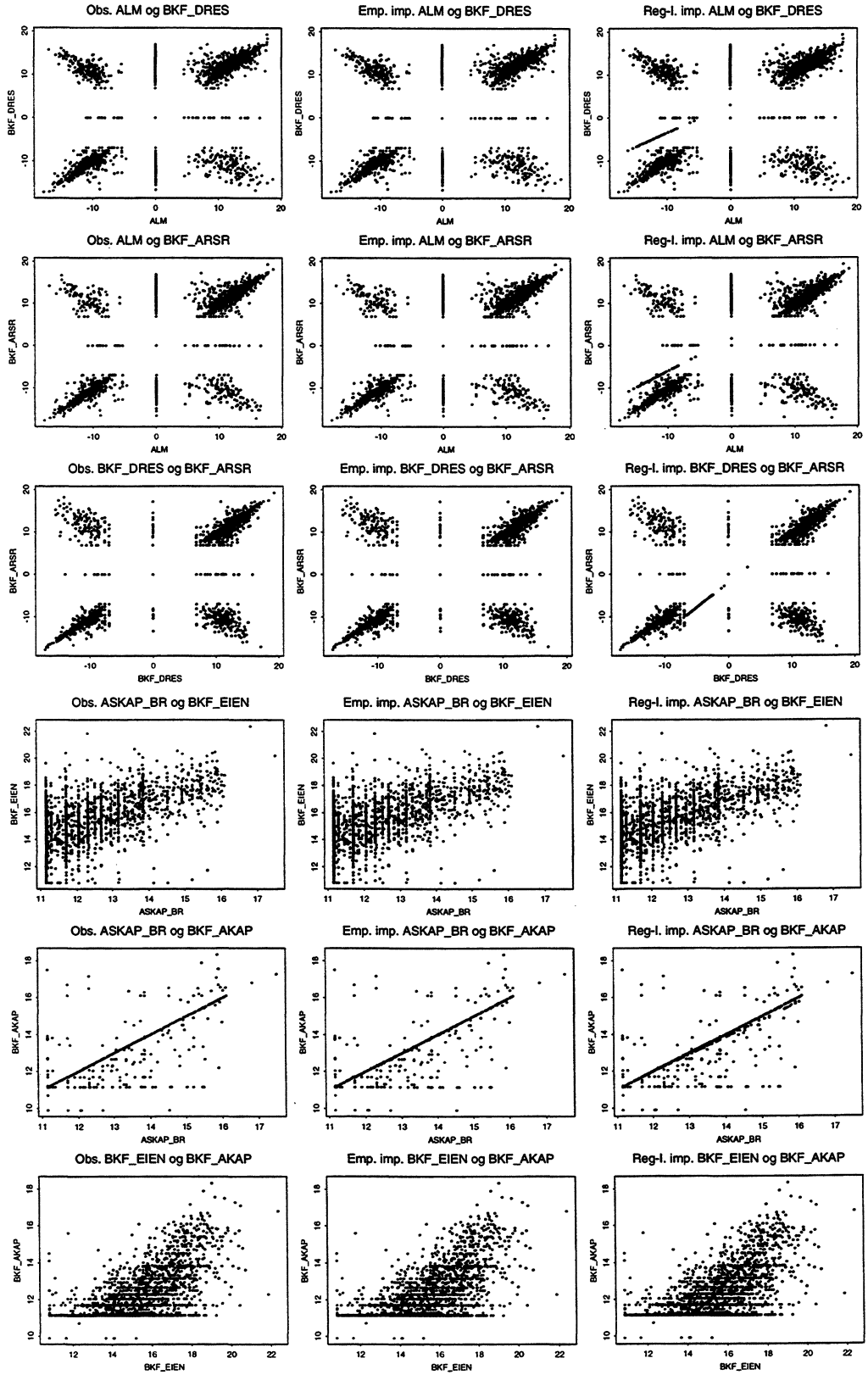
Sampel-	Transformerte data						
varians	Emp-I	Emp-II	Reg-I	Reg-II	Fasiten	Met-I	Met-0
BKF-DRI	19.117	19.126	18.617	18.968	19.309	19.613	27.828
Relat. Diff.	-0.009	-0.008	-0.035	-0.017	0	0.016	0.441
BKF-DRES	114.8	114.833	110.85	112.56	114.669	114.15	109.558
Relat. Diff.	0.001	0.001	-0.032	-0.017	0	-0.004	-0.044
BKF-ARSR	117.788	117.733	114.635	116.723	117.533	116.955	111.772
Relat. Diff.	0.002	0.002	-0.024	-0.006	0	-0.004	-0.048
BKF-EIEN	3.2	3.194	3.151	3.246	3.261	3.244	3.949
Relat. Diff.	-0.018	-0.02	-0.033	-0.004	0	-0.004	0.211
BKF-AKAP	1.445	1.446	1.435	1.465	1.449	1.419	1.547
Relat. Diff.	-0.002	-0.001	-0.009	0.011	0	-0.02	0.068

Standard-	Opprinnelige skala						
Avvik	Emp-I	Emp-II	Reg-I	Reg-II	Fasiten	Met-I	Met-0
BKF-DRI	39274593	39289337	39220773	39319840	40919497	40240547	39244827
Relat. Diff.	-0.039	-0.039	-0.041	-0.038	0	-0.016	-0.04
BKF-DRES	5192506	5180394	5159438	5178920	5190611	5300829	5159812
Relat. Diff.	0.000	-0.001	-0.005	-0.001	0	0.021	-0.005
BKF-ARSR	5105108	5092587	5065143	5104402	5158275	5204315	5065090
Ralat. Diff.	-0.009	-0.012	-0.017	-0.009	0	0.009	-0.017
BKF-EIEN	58318150	58314018	58314883	58528561	58589940	59882295	58304662
Ralat. Diff.	-0.004	-0.004	-0.004	-0.000	0	0.022	-0.004
BKF-AKAP	2734361	2734473	2727197	2742548	2739425	2789793	2717457
Ralat. Diff.	-0.001	-0.001	-0.003	0.001	0	0.018	-0.007



4.3 Resultater (II)

Vi skal se litt på den simultane variasjonen i data. (Empirisk imputering som presenteres i dette avsnittet er generert med vindusbredden $\alpha = 0.05$.)



Man kan se i de fleste tilfellene at de imputerte verdiene under Reg-I skiller seg ut fra de observerte data — de ligger på en linje. Derimot er det umulig å identifisere på samme måten de imputerte verdiene i empirisk imputering. Til slutt er det i tabellen nedenfor listet ut parvise korrelasjoner, observerte mot imputerte, på transformerte data.

	Observert korrelasjon (3203 foretak)					
	BKF-DRES	BKF-ARSR	BKF-EIEN	BKF-AKAP	ASKAP-BR	ALM
BKF-DRI	0.43	0.27	0.51	0.19	0.18	0.24
BKF-DRES		0.76	0.33	0.06	0.06	0.62
BKF-ARSR			0.31	0.11	0.11	0.73
BKF-EIEN				0.59	0.58	0.34
BKF-AKAP					0.93	0.09
ASKAP-BR						0.10
	Imputert korrelasjon (3383 foretak)					
	BKF-DRES	BKF-ARSR	BKF-EIEN	BKF-AKAP	ASKAP-BR	ALM
BKF-DRI	0.41	0.26	0.51	0.20	0.19	0.23
BKF-DRES		0.77	0.31	0.06	0.06	0.64
BKF-ARSR			0.30	0.10	0.10	0.74
BKF-EIEN				0.60	0.59	0.32
BKF-AKAP					0.93	0.08
ASKAP-BR						0.09

4.4 Resultater (III)

Dagens opplegg for vekting for IFS kalles kalibrering. Vi har derfor undersøkt hvordan imputering fungerer i denne forbindelsen på følgende enkle måte: (1) sett initiale vekter til alle 3383 foretak til $N/n = 97049/3383$ der $N = 97049$ er antall foretak i populasjonen; (2) kalibrer mot populasjonstotaler (N, BKF-DRI, BKF-DRES, BKF-ARSR, BKF-EIEN, BKF-AKAP, ASKAP-BR, ALM), basert på (a) fasitten, (b) empirisk imputering (Emp-I), og (c) betinget forventningsimputering (Reg-I); (3) til slutt kalibrer mot kun (N, ALM) som en forenklet utgave av dagens metode der man ikke trenger å imputere før kalibrering.

For å vurdere mer detaljert hvor nær et vektsett, betegnet med $\{w_i\}$ for $i = 1, ..., 3383$, ligger i forhold til et annet, betegnet med $\{a_i\}$, beregner vi to indekser, nemlig

$$D(w,a) = \sum_{i=1}^{3383} \frac{1}{2}(\frac{w_i}{a_i} - 1)^2 \qquad R(w,a) = \sum_{i=1}^{3383} (\frac{w_i}{a_i} - 1)/3383.$$

Mens D er en metrikkavstand, står R for gjennomsnittlig relativt avvik fra $\{w_i\}$ til $\{a_i\}$. Vi finner følgende tall:

$$\begin{aligned} D(\text{Reg-I, Fasiten}) &= 3331 & R(\text{Reg-I, Fasiten}) &= -0.026 \\ D(\text{Emp-I, Fasiten}) &= 1507 & R(\text{Emp-I, Fasiten}) &= -0.018 \\ D(\text{Dagens, Fasiten}) &= 4778262 & R(\text{Dagens, Fasiten}) &= 0.907. \end{aligned}$$

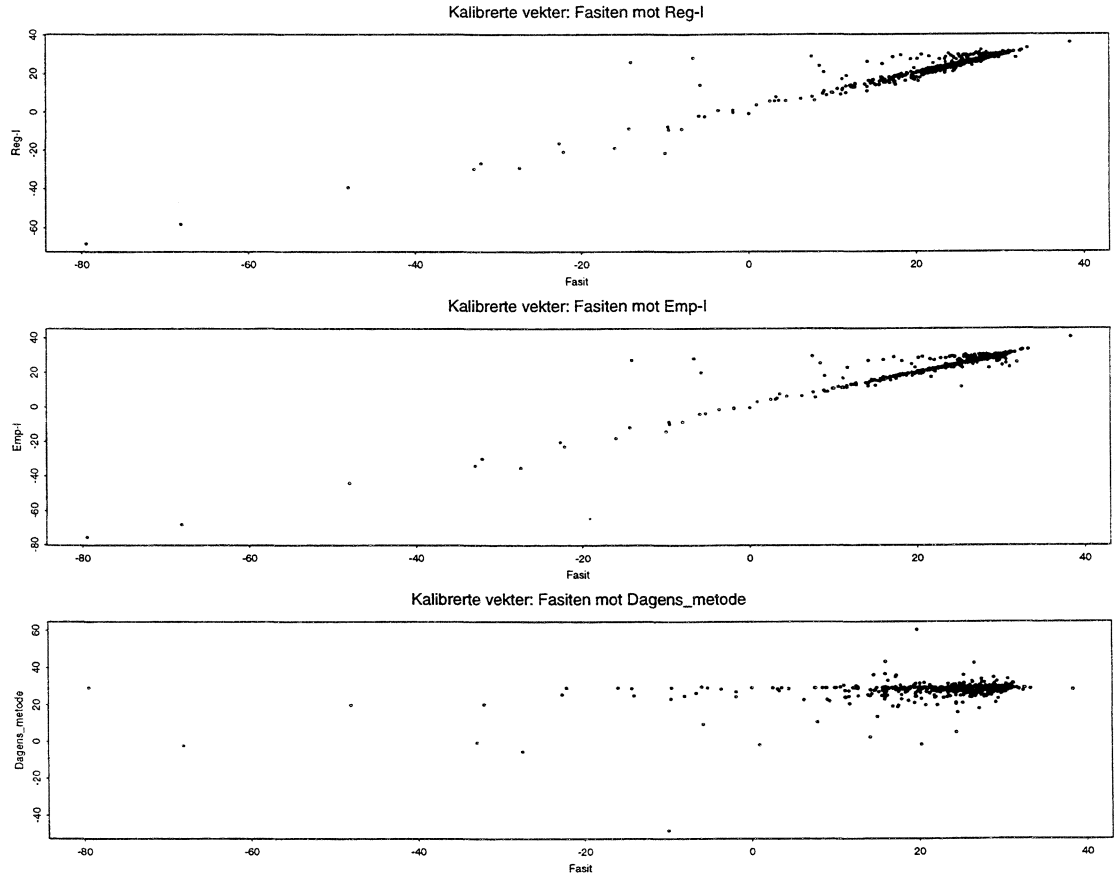
I tillegg kan man undersøke skjevheten til dagens metode ved å sjekke estimater for kjente registertotaler basert på vektene kalibrert etter dagens metode. (Man får igjen eksakt alle disse totalene ved å bruke vektene under Emp-I eller Reg-I.)

	N	BKF-DRI	BKF-DRES	BKF-ARSR
Total	97049	547280486000	30999448000	21788654000
Estimat	97049	1103295579592	43694638663	19945132811
Relativt avvik. (%)	0.0	101.6	41.0	-8.5
	BKF-EIEN	BKF-AKAP	ASKAP-BR	ALM
Total	435225975000	29004941000	30557788679	22776845777
Estimat	730938460745	53376812309	42064145903	22776845777
Relativt avvik (%)	67.9	84.0	37.7	-0.0

Kommentar 23 *I realiteten er nok dagens vekting mer komplisert enn som så. Men det er liten tvil om at kalibrering mot ALM har omt. ingen effekt på andre register-variablene.*

Kommentar 24 *En annen måte å se dette på er å notere at to blant de 7 variablene som er studert her finnes i hele populasjon. Dersom man ikke tar i bruk de 5 BKF-variablene pga. c.a.* 10% *frafall der, kaster man vekk minst* $5/7 \approx 71\%$ *informasjon.*

Til slutt har vi plottet de kalibrerte vektene ved Reg-I, Emp-I og dagens metode mot fasitten:



5 Diskusjon

Oppsummering:

- Man bør satse på imputering selv når partielt frafall er tilfeldig, fordi det ligger informasjon i de observerte verdiene hos enhetene med partielt frafall, som ville gå tapt dersom man bare benytter seg av den komplette delen av data.

- Betinget forventingsimputering kan rette opp mye for enkelte marginale nivåer gitt passende modell, men vil generelt underslå variasjonen i data. I tillegg kan regresjonsmodellen som ble brukt i dette studiet være følsom overfor valget av skala. Det er også tegn på at regresjonen (Reg-II) kan være ustabil pga. ekstreme verdier.

- Muligens kan empirisk imputering lettest forstås som kanonisert nærmeste nabo (imputering). Begrepet nærmeste nabo er klargjort gjennom betinget fordeling. Metoden sikter direkte på den multivariate empiriske fordelingen, som inneholder all informasjon i data. Den har vist seg å være bedre enn komponentvis hot-deck, både simultant og marginalt, ved å følge den betingete fremgangsmåten. Den er bedre enn betinget forventingsimputering til å bevare (ko-)variasjon i data, samtidig som den er mer robust overfor valget av skala.

  Kommentar 25 *Grunnen til at empirisk imputering er mindre følsom overfor skala skyldes antageligvis at den empiriske tetthetsfordelingen er skalainvariant. I motsetning er parametrene i en regresjonsmodell direkte påvirket av verdiene, og derfor skala, til observersjonene.*

Forslag til videreutvikling:

- Generelle studier om valget av vindu med betinging på flere variabler samtidig. Dette er antakelig den viktigste metodiske problemstillingen ved empirisk imputering. Valget her bestemmer hvilke betingete empiriske fordelinger som skal danne grunnlaget for imputering, og på den måten avgjøre resultatet.

  Kommentar 26 *Resultatene i dette studiet viser at metoden kan være rimelig robust overfor valget av vindusbredden i tilfellet betinging med hensyn på en variabel på passende skala.*

- Alternative, kanskje glattere eller/og mer stabile, ikke-parametriske betingete fordelinger enn den empiriske, som muligens fungerer bedre i tilfeller med et lite antall obervasjoner, til tross for at vi ikke har funnet noe behov i dette studiet.

- Effektive iterative algoritmer, som ligner på Gibbs-sampler, for å takle den simultane empiriske fordelingen direkte. (Vi har brukt diverse forenklinger her som f.eks. gruppering av variablene, osv.) Problemstillingen vil være høyst aktuell i tilfellet med mer kompliserte frafallsmønstre, der forskjellige variabler faller fra på forskjellige enheter og, kanskje, ingen av variablene er tilgjengelig overalt.

  Eksempel 7 *Som et eksempel anta bivariate observasjoner*

$$(Y_1, Y_2) = (1, -),\ (-, 1),\ (1, 1),\ (2, 1),\ (1, 2),\ (2, -).$$

  Uten først å har imputert for den ene, så kan man ikke imputere for den andre basert på betinget fordeling.

De sist utgitte publikasjonene i serien Notater

98/16 A.A. Ritland: Livsstil, seksualitet og helse: En spørreskjemaundersøkelse: Dokumentasjonsrapport. 13s.

98/17 A.A. Ritland: Seksualitet og helse: En spørreskjemaundersøkelse: Dokumentasjonsrapport. 24s.

98/18 H.M. Teigum: Kostholdsundersøkelsen 1997: Dokumentasjonsrapport. 38s.

98/19 C. Hendriks: FoB2000: Rapport fra seminar 18. mars 1998 om kjennemerker i boligtellingen. 41s.

98/20 D.Q. Pham: Sesongjustering av tidsserier i Statistisk sentralbyrå: En sammenligning mellom X11 ARIMA og X12 ARIMA. 85s.

98/21 F. Bendiksen og K.-A. Hovland: Foreldrebetalingsundersøkelse: Rapport om betalingen for heldagsopphold i kommunale og private barnehager. 1. halvår 1998. 36s.

98/22 L. Lindholt: Dynamiske oljemodeller: Intertemporal optimering og adferdssimulering. 55s.

98/23 T.N. Evensen: Nasjonalregnskap: Beregning av post- og distribusjonsvirksomhet. 23s.

98/24 P.M. Holt, L. Haugen og P.E. Gjedtjernet: Skattestatistikk. Etterskuddspliktige 1995 og 1996: Dokumentasjon. 36s.

98/25 Regionale inndelinger: En oversikt over standarder i norsk offisiell statistikk. 130s.

98/26 L. Rogstad: FoB 2000. Geografisk informasjon i Folke- og boligtellingen år 2000: En oversikt over sentrale regionale kjennemerker og inndelinger. 36s.

98/27 L. Rogstad: FoB2000: Rapport fra seminar 12. februar 1998 om geografisk informasjon i Folke- og boligtellingen år 2000. 46s.

98/28 E. Midtlyng: Dokumentasjonsrapport AKU 1996. 41s.

98/29 G. Haakonsen, K. Rypdal og B. Tornsjø: Utslippsfaktorer for lokale utslipp - PAH, partikler og NMVOC. 74s.

98/30 FoB2000. Folke- og boligtellingen år 2000: Høringsnotat om innhold. 49s.

98/31 G. Dahl og J. Folkedal: FD - Trygd. Dokumentasjonsrapport: Stønader til enslig forsørger, 1992-1993. 34s.

98/32 K. Bjønnes og J. Johansen: FD - Trygd. Dokumentasjonsrapport: Attføringspenger, 1992-1993. 108s.

98/33 O. Skorge: Forsknings- og utviklingsvirksomhet (FoU) 1995: Dokumentasjon av FoU-undersøkelsen 1995. 30s.

98/34 A. Sundvoll og H.M. Teigum: Samordnet levekårsundersøkelse 1997 - tverrsnittsundersøkelsen: Dokumentasjonsrapport. 130s.

98/35 K. J. Einarsen, A. B. Skara og C. Strand: Faktaark for FylkesKOSTRA-utdanning. 1. tertial 1998. Sør-Trøndelag fylkeskommune: Nøkkeltall med indikatorer for Prioriteringer, Dekningsgrad, Produktivitet. 39s.

98/36 P. Bakken og J.A. Osnes: Kvartalsvis ordrestatistikk. 53s.

98/39 I. Melby og R. Aaberge: Sammenligning og fordeling av husholdsinntekt blant barn og unge. 31s.

98/40 A.A. Ritland: Evaluering av Reform 94. En spørreskjemaundersøkelse: Dokumentasjonsrapport. 43s.

98/41 D. Roll-Hansen, L. Solheim og L.C. Zhang: Kopiering ved universiteter og høgskoler. 88s.

98/42 M.V. Dysterud og P. Schøning: Etterprøvbare miljømål for byer og tettsteder: Et metodeprosjekt for utvikling og prøving av miljøindikatorer. 40s.

98/43 J. Epland: Inntekt etter skatt: Revisjon av inntektsregnskapet i inntekts- og formuesundersøkelsen for husholdninger. 40s.

98/45 L. Aaram og Ø. Skullerud: Statistikk over emballasjeavfall: Utprøving av metode og foreløpige resultater. 32s.

Notater

PORTO BETALT
VED
INNLEVERINGA
P.P.
NORGE/NOREG

Tillatelse nr.
159 000/502

B *Returadresse:*
Statistisk sentralbyrå
Postboks 8131 Dep.
N-0033 Oslo

Statistisk sentralbyrå

Oslo:
Postboks 8131 Dep.
0033 Oslo

Telefon: 22 86 45 00
Telefaks: 22 86 49 73

Kongsvinger:
Postboks 1260
2201 Kongsvinger

Telefon: 62 88 50 00
Telefaks: 62 88 50 30

ISSN 0806-3745

Statistisk sentralbyrå
Statistics Norway