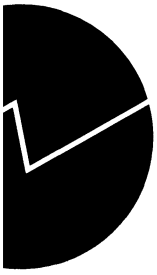
Statistics Norway



Statistisk sentralbyrå

Roger Jensen

Beregning av usikkerhet for boligprisindeksene på grunn av frafall

Roger Jensen

Beregning av usikkerhet for boligprisindeksene på grunn av frafall

Statistisk sentralbyrå • Statistics Norway
Oslo–Kongsvinger

Rapporter

I denne serien publiseres statistiske analyser, metode- og modellbeskrivelser fra de enkelte forsknings- og statistikkområder. Også resultater av ulike enkeltundersøkelser publiseres her, oftest med utfyllende kommentarer og analyser.

Reports

This series contains statistical analyses and method and model descriptions from the different research and statistics areas. Results of various single surveys are also published here, usually with supplementary comments and analyses.

ISBN 82-537-4669-5
ISSN 0806-2056

Emnegruppe
Emnegruppe
08.02 Prisindekser

Emneord
Bruktbolig
Enebolig
Frafallsfeil
Indekser
Indeksverdier

Design: Enzo Finger Design
Trykk: Statistisk sentralbyrå

Standardtegn i tabeller	Symbols in tables	Symbol
Tall kan ikke forekomme	Category not applicable	.
Oppgave mangler	Data not available	..
Oppgave mangler foreløpig	Data not yet available	...
Tall kan ikke offentliggjøres	Not for publication	:
Null	Nil	-
Mindre enn 0,5 av den brukte enheten	Less than 0.5 of unit employed	0
Mindre enn 0,05 av den brukte enheten	Less than 0.05 of unit employed	0,0
Foreløpige tall	Provisional or preliminary figure	*
Brudd i den loddrette serien	Break in the homogeneity of a vertical series	—
Brudd i den vannrette serien	Break in the homogeneity of a horizontal series	|
Rettet siden forrige utgave	Revised since the previous issue	r

Sammendrag

Roger Jensen

Beregning av usikkerhet for boligprisindeksene på grunn av frafall

Rapporter 99/4 • Statistisk sentralbyrå 1999

De estimerte standardavvikene beregnet i denne rapporten er et ledd i arbeidet med kvalitetssikring av Statistisk sentralbyrås statistikker. Utgangspunktet er at ingen estimerte indekser er eksakte. Ved gjentak av hele produksjonsprosedyren vil resultatene som oftest bli ulike fra gang til gang. Det er ulike årsaker til dette, for eksempel at utvalget vil være ulikt sammensatt, feil i eventuelle registre, feil/endringer i innrapporteringer fra oppgavegivere, eller feil i Statistisk sentralbyrås registrering av innkomne data.

I skjemabaserte undersøkelser vil det som oftest være et visst frafall fordi ikke alle oppgavegiverne returnerer sine skjema i korrekt utfylt stand. De variasjoner i indeksestimatene en får på grunn av dette, kalles frafallsfeil.

Rapporten gir en beskrivelse av tre metoder for estimering av standardavviket på grunn av frafall. Metodene som er benyttet er bootstrapping, jackknife og kryssvalidering. Alle metodene er generelle, og kan også benyttes på andre typer estimatorer enn indekser. Resultatene viser at bootstrapping er den beste metoden å bruke for å estimere standardavviket for boligprisindeksene.

I prisindeks for nye eneboliger ligger det estimerte standardavviket på grunn av frafall mellom 0,5 og 0,7 prosentpoeng, mens det estimerte standardavviket for totalindeksen i prisindeks for bruktbolig ligger mellom 0,3 og 0,4 prosentpoeng. Publisert totalindeks i prisindeks for bruktbolig 1. kvartal 1997 var 132,2. Med et estimert standardavvik på 0,4 prosentpoeng kan vi, dersom vi antar at det er ingen andre feilkilder enn frafall, være 90 prosent sikre på at den virkelig indeksverdien dette kvartalet ligger mellom 131,6 og 132,8.

Emneord: Bruktbolig, enebolig, frafallsfeil, indekser, indeksverdier.

Innhold

1. Innledning **7**

2. Tolkning av standardavviket **8**

3. Beregningsmetode for prisindeksene **9**
3.1. Konstruksjon av prisindeks for nye eneboliger 9
3.2. Konstruksjon av prisindeks for bruktbolig 10

4. Varians i indeksene på grunn av frafall **11**

5. Stokastisk simulering for beregning av varians **12**
5.1. Bootstrapping 13
5.1.1. Parametrisk bootstrap 13
5.1.2. Vanlig (ikke-parametrisk) bootstrap 13
5.2. Jackknife 13
5.3. Kryssvalidering 14
5.3.1. Antall grupper 14

6. Intervallestimering **15**
6.1. Standardintervallet 15
6.2. Bootstrap konfidensintervall 15
6.2.1. Persentilmetoden 15
6.2.2. BC-intervallet 15

7. Numeriske resultater **17**

8. Sammenligning av modellene **22**

9. Oppsummering og konklusjoner **23**

Referanser **24**

De sist utgitte publikasjonene i serien Rapporter **25**

1. Innledning

Ved beregning av prisindeks for nye eneboliger og prisindeks for bruktbolig ligger det hvert kvartal et visst antall observasjoner av henholdsvis fullførte nye eneboliger og omsatte brukte boliger til grunn. Ingen av disse estimerte indeksverdiene er imidlertid helt eksakte. I beregning av indeksene benyttes i tillegg til opplysninger om prisen en rekke tilleggsopplysninger om den enkelte bolig. Dette er ulike opplysninger om boligens standard; areal, antall bad, antall WC osv. Tilleggsopplysningene brukes til å justere indeksene for kvalitetsendringer i boligmassen fra kvartal til kvartal. En del av disse opplysningene hentes direkte fra Grunneiendoms- ,Adresse- og Bygningsregisteret (GAB-registeret). I tillegg sendes et spørreskjema til eierne av boligene for å hente inn tilleggsopplysninger om boligene.

En kilde til feil i indeksene er ulike typer frafall i undersøkelsene. Noen spørreskjema får vi ikke sendt ut fordi vi ikke finner adressen til vedkommende som skulle hatt spørreskjemaet. En del oppgavegivere returnerer ikke sine skjema. I tillegg kan noen skjema være ufullstendig utfylt, slik at de må forkastes. Den estimerte indeksverdien i et kvartal vil variere alt etter hvilke av boligene det er som ligger til grunn for beregningene. I dette notatet er indeksenes standard-avvik beregnet på grunnlag av at en antar at frafallet er tilfeldig. Standardavviket vil i tillegg til selve frafallet også være påvirket av antall observasjoner som ligger til grunn for beregningene og prisvariasjonen mellom observasjonene.

2. Tolkning av standardavviket

Det beregnede standardavviket til en indeks I som er beregnet i denne rapporten kan si noe om usikkerheten vi har i de estimerte indeksverdiene når en antar at frafallet er tilfeldig. Dess mindre standardavviket er, dess mer nøyaktig er den estimerte indeksverdien.

Vi antar videre i denne rapporten at fordelingen til indeksestimatoren $\hat{I}^{k,T}$ for kvartal k i år T er normalfordelt. For en estimator $\hat{I}^{k,T}$ med standardavvik std($\hat{I}^{k,T}$) er vi 90 prosent sikre på at den ukjente indeksverdien I ligger innenfor intervallet $\pm 1{,}6 \cdot$ std($\hat{I}^{k,T}$). Dette intervallet vil i 90 av 100 tilfeller dekke den ukjente indeksverdien. Denne metoden for konstruksjon av konfidensintervall kalles standardintervallet og er også beskrevet i kapittel 7.1. Et par andre aktuelle teknikker for konstruksjon av konfidensintervaller er også beskrevet i kapittel 7.

Indekser blir brukt til å måle endring mellom to perioder. Endringen fra et kvartal til neste kan skrives som

$$\Delta^{k,T} = I^{k+1,T} - I^{k,T}$$

hvor $\Delta^{k,T}$ er endringen i indeksverdien fra kvartal k til kvartal $k+1$ og $I^{k+1,T}$ og $I^{k,T}$ er indeksverdier i kvartal $k+1$ og kvartal k i år T. Et standardavvik for estimatoren $\hat{\Delta}^{k,T}$ kan si noe om usikkerheten i endringen i indeksverdien mellom disse to kvartalene. Dess lavere standardavviket er, dess mer nøyaktig vil endringstallet være.

Vi antar også her at fordelingen til estimatoren $\hat{\Delta}^{k,T}$ er normalfordelt. For en estimator $\hat{\Delta}^{k,T}$ med standardavvik std($\hat{\Delta}^{k,T}$) er vi 90 prosent sikre på at den ukjente endringen $\Delta^{k,T}$ ligger innenfor intervallet $\pm 1{,}6 \cdot std(\hat{\Delta}^{k,T})$. Dersom estimatoren for eksempel har et standardavvik på 2 prosentpoeng, betyr dette at intervallet $\left[\hat{\Delta}^{k,T} - 1{,}6 \cdot 2, \hat{\Delta} + 1{,}6 \cdot 2\right]$ er et 90 prosent konfidensintervall for $\Delta^{k,T}$.

For å avgjøre om en endring er reell, det vil si at indeksverdiene i de to periodene vi ser på er ulike, må vi først sette et signifikansnivå. Med et signifikansnivå α er vi (1-α) prosent sikre på at den konklusjonen vi trekker er riktig. For et signifikansnivå α konkluderer vi med at indeksverdiene i to perioder er ulike dersom et (1-α) konfidensintervall for endring ikke dekker verdien 0. Konkrete eksempler på konfidensintervall for de estimerte indeksene og test på om endringene er reelle er gitt i kapittel 9.

3. Beregningsmetode for prisindeksene

Både nye og brukte boliger varierer betydelig både i størrelse og utforming. Dette forholdet gjør at beregning av prisindeks for bruktbolig og prisindeks for nye eneboliger blir komplisert. For å få fram en korrekt prisendring mellom to kvartaler, er det nødvendig å benytte spesielle analysemetoder som justerer for prisendringer som skyldes kvalitetsforskjeller. Med kvalitetsforskjeller mener vi ulik standard (areal, antall WC, antall bad, type ventilasjon o.l.). Ved hjelp av regresjonsanalyse kan en kartlegge og prise de ulike kvalitetsegenskapene.

Generelt prøver man i en regresjonsanalyse å beregne hvordan en variabel endres (avhengig variabel) når en eller flere andre variabler endres (forklaringsvariabler). Ved beregning av prisindeks for nye eneboliger er kvadratmeterprisen for eneboligene den avhengige variabelen. Dette fordi man her er ute etter å måle utviklingen i prisen pr. kvadratmeter byggherre/kjøper må betale for en ny enebolig. I prisindeks for bruktbolig er derimot prisen for boligen den avhengige variabelen. Her ønsker man å måle utviklingen i prisen kjøper må betale for en brukt bolig. Det benyttes i begge tilfeller en lineær regresjonsmodell estimert ved minste kvadraters metode.

3.1. Konstruksjon av prisindeks for nye eneboliger

I en regresjonsanalyse med flere forklaringsvariable kan sammenhengen uttrykkes som følger:

$$(3.1)\quad Y_{it} = a_t + \sum_{k=1}^{K} b_{kt} \cdot X_{ikt} + \varepsilon_{it}$$

der:

Y_{it} = kvadratmeterprisen på bolig i på tidspunkt t
a_t = et fast beløp per bolig på tidspunkt t, uavhengig av boligens øvrige egenskaper
b_{kt} = enhetspris på den k-te kvalitetsegenskap på tidspunkt t
X_{ikt} = kvantum av den k-te kvalitetsegenskap for i-te bolig på tidspunkt t
ε_{it} = restledd som ivaretar tilfeldig variasjon. Restleddet antas å ha konstant varians med forventningsverdi lik 0.

Prisindeks for nye eneboliger måler prisutviklingen per kvadratmeter. Indeksuttrykket ved bruk av flere forklaringsvariable kan uttrykkes ved følgende indeksformel:

$$(3.2)\quad I^{P}{}_{0t} = \frac{a_t + \sum_{k=1}^{K} b_{kt}\overline{X}_{kt}}{a_0 + \sum_{k=1}^{K} b_{k0}\overline{X}_{kt}} \cdot 100$$

Uttrykket i (3.2) er en Paascheprisindeks. Gjennomsnittsprisen per kvadratmeter i hver periode kan skrives som en funksjon av de gjennomsnittlige verdier i settet av forklaringsvariable. Dette kan skrives som

$$(3.3)\quad \overline{Y}_0 = a_0 + \sum_{k=1}^{K} b_{k0}\overline{X}_{k0}$$

Gjennomsnittlig kvadratmeterpris i periode 0

$$(3.4)\quad \overline{Y}_t = a_t + \sum_{k=1}^{K} b_{kt}\overline{X}_{kt}$$

Gjennomsnittlig kvadratmeterpris i periode t

Ved å løse uttrykkene i (3.3) og (3.4) på henholdsvis a_0 og a_t og sette inn i indeksuttrykket (3.2) får vi følgende uttrykk for prisindeksen:

$$(3.5)\quad I_{0t}{}^{P} = \frac{\overline{Y}_t}{\overline{Y}_0 + \sum_{k=1}^{K} b_{k0}(\overline{X}_{kt} - \overline{X}_{k0})} \cdot 100$$

Uttrykket i teller er gjennomsnittlig kvadratmeterpris observert i periode t, som er sammenligningsperioden. Første ledd i nevner er observert kvadratmeterpris i periode 0, som er basisperioden. I annet ledd i nevner

prises differansen i det gjennomsnittlige kvantum av kvalitetsegenskapene i de to periodene. Uttrykket i nevner er en beregnet kvadratmeterpris for gjennomsnittsboligen i periode t dersom den var bygd i periode 0. Indeksen uttrykker derfor prisutviklingen i kvadratmeterpriser for boliger av lik standard.

Det er viktig å merke seg at i en Paascheindeks er det kun nødvendig med regresjonsberegninger for basisperioden, slik som det også framgår av indeksuttrykket i (3.5). Dette har flere fordeler. For det første blir selve arbeidet med indeksen enklere når man kun trenger å beregne regresjonsligningen for basis. I prisindeks for nye eneboliger er basis ett år, mens sammenligningsperiodene er kvartaler. Antall observasjoner i kvartalene kan være så små at usikkerheten ved regresjonsanalyser kan bli vesentlig større enn for en regresjonsanalyse med observasjoner fra en lengre periode. For å få godt estimerte regresjonskoeffisienter blir regresjonsligningen her beregnet med data for de to siste år. Indeksen er en kjedeindeks med årlige lenker, det vil si at man bytter basis hvert år. Indeksen tar derfor hensyn til at forholdet mellom priser og kvalitative egenskaper ved boligene skifter karakter over tid.

3.2. Konstruksjon av prisindeks for bruktbolig

I prisindeks for bruktbolig er det den naturlige logaritmen til prisen som er avhengig variabel. Hvis vi kaller de numeriske kvalitetsvariablene $X_1,...,X_K$ og dummyvariablene $Z_1,...,Z_L$ kan regresjonsmodellen skrives på formen

$$(3.6)\quad \ln Y_{it} = a_t + \sum_{k=1}^{K} b_{kt} \ln X_{ikt} + \sum_{l=1}^{L} c_{lt} Z_{ilt} + \varepsilon_{it}$$

Det kan vises at vi her får følgende uttrykk for prisindeksen

$$(3.7)\quad I_{0t}{}^{P} = \frac{\exp(\ln \overline{Y}_t - \sum_{k=1}^{K} b_{k0} \cdot \ln \overline{X}_{kt} - \sum_{l=1}^{L} c_{l0} \cdot \overline{Z}_{lt})}{\exp(\ln \overline{Y}_0 - \sum_{k=1}^{K} b_{k0} \cdot \ln \overline{X}_{k0} - \sum_{l=1}^{L} c_{l0} \cdot \overline{Z}_{l0})}$$

Indeksuttrykket i (3.7) kan, ved å trekke sammen leddene i teller og nevner, også skrives på samme form som indeksuttrykket i (3.5).

Her beregnes også koeffisientene i regresjonsligningen på grunnlag av data for de to siste år, mens basis er siste år.

4. Varians i indeksene på grunn av frafall

Ved beregning av prisindeks for nye eneboliger og prisindeks for bruktbolig ligger det hvert kvartal et visst antall observasjoner av henholdsvis nye eneboliger og omsatte bruktboliger til grunn. I beregning av indeksene benyttes i tillegg til opplysninger om prisen en rekke tilleggsopplysninger om den enkelte bolig. Dette er ulike opplysninger om boligens standard: areal, antall bad, antall WC osv. En god del av disse tilleggsopplysningene hentes inn via et spørreskjema som sendes postalt til kjøpere og eiere av de aktuelle boligene.

En kilde til feil i indeksen er ulike typer frafall. Noen skjema får vi ikke sendt ut fordi vi ikke har funnet adressen til vedkommende som skulle hatt skjemaet. En del oppgavegivere returnerer ikke sine skjema. I tillegg kan noen skjema være ufullstendig utfylt, slik at de må forkastes. Dette frafallet gjør at vi får en usikkerhet i indeksen. Det er indeksens varians som følge av frafall vi her ønsker å måle.

Det datagrunnlaget som i kvartal t ligger til grunn for beregning av indeksene kan skjematisk beskrives slik:

Bolig	Pris	Kvalitetsvariabler (areal, antall bad, antall WC osv.)				
1	Y_{1t}	X_{11t}	X_{12t}	.	.	X_{1Kt}
2	Y_{2t}	X_{21i}	X_{22t}	.	.	X_{2Kt}
.	.	.	.	.		.
n	Y_{nt}	X_{n1i}	X_{n2t}	.	.	X_{nKt}
n+1	Y_{n+1t}	X_{n+11i}	X_{n+12t}	.	.	X_{n+1Kt}
.	.	.	.	.	.	.
N	Y_{Nt}	X_{N1t}	X_{N2t}	.	.	X_{NKt}

I prisindeks for nye eneboliger vil det hvert kvartal være N fullførte eneboliger. Av alle de N eierne av nye eneboliger som mottar skjema er det bare *n* av disse som returnerer skjema i korrekt utfylt stand. På bakgrunn av de *n* eneboligene vi i kvartal t legger til grunn for beregningene får vi beregnet følgende størrelser:

$$\overline{Y}_t,\ \overline{X}_{1t},\ \overline{X}_{2t},\ \ldots\ \overline{X}_{Kt}$$

Det vil dermed være usikkerhet i alle disse størrelsene på grunn av frafall.

Tilsvarende vil det være usikkerhet på grunn av frafall i de størrelsene som beregnes fra basis, der basis er data fra året før:

$$\overline{X}_{10},\quad \overline{X}_{20},\ \ldots\ \overline{X}_{K0}$$

Til slutt er det usikkerhet på grunn av frafall i de estimerte regresjonskoeffisientene b_{10}, b_{20},... ,b_{K0}. Disse er estimert på grunnlag av data fra de to foregående år.

Et helt tilsvarende resonnement får vi for beregning av prisindeks for bruktbolig.

Problemet med estimering av variansen til indeksene kan generelt beskrives slik:

Et utvalg $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ fra en ukjent sannsynlighetsfordeling **F** er observert, og vi ønsker å estimere en indeks θ på grunnlag av **X**. Vi gjør dette ved å estimere $\hat{\theta} = I_n(\mathbf{X})$ fra **X**. For å kunne si noe om usikkerheten til indeksen, ønsker vi å kjenne variansen $Var_\theta I_n$, eller eventuelt et estimat ($\hat{Var}_\theta I_n$) for denne.

I beregning av prisindeks for bruktbolig og prisindeks for nye eneboliger består ikke **X** bare av observasjonene i det aktuelle kvartalet, men i kombinasjon med observasjonene fra de foregående to år. Observasjonene fra foregående år benyttes til basis og observasjonene for de to foregående år til beregning av regresjonskoeffisientene.

I de neste kapitlene er det beskrevet tre metoder for estimering av variansen til indekser. Metodene er generelle, og kan også brukes på andre typer estimatorer enn indekser.

5. Stokastisk simulering for beregning av varians

Den første av de tre metodene for stokastisk simulering som er beskrevet i dette kapitlet kalles bootstrapping. Metoden er blant annet beskrevet i Efron og Tibshirani (1993) og Lindqvist (1996). Den andre metoden, som kalles jackknife, er beskrevet i Efron og Tibshirani (1993). Den siste metoden kalles kryssvalidering og er beskrevet i Lillegård (1994).

Vi ser først på følgende ideelle situasjon. Anta at vi kjenner den sanne verdi av indeksen, θ_0 av θ. Da har man fullstendig kjennskap til sannsynlighetsfordelingene til observasjonene $\mathbf{X}_i$, den er gitt ved sannsynlighetsfordelingen $\mathbf{F}$. I prinsippet kan man da regne ut $Var_{\theta_0} I_n$ og dersom vi også kan gjennomføre det i praksis, er vi ferdige. Anta imidlertid at dette ikke lar seg løse analytisk. Siden fordelingen for $\mathbf{X}$-ene er kjent, kan variansen da approksimeres med så stor nøyaktighet en ønsker ved hjelp av stokastisk simulering, også kalt Monte Carlo-simulering.

Ideen ved stokastisk simulering er at man ved hjelp av en random generator simulerer uavhengige («tenkte») realisasjoner av observasjonene ved hjelp av en datamaskin. Gitt en slik mulighet for å generere observasjoner av $\mathbf{X}$, kan vi nå også generere realisasjoner av indeksen, $I_n(\mathbf{X})$. Dette gjøres først ved å generere $\mathbf{X} = (\mathbf{X}_1,...,\mathbf{X}_n)$ ved å trekke n-uavhengige verdier for $\mathbf{X}$, og så regne ut $I_n(\mathbf{X})$. Et anslag for Var_{θ_0} får vi ved på denne måten å generere et (stort) antall k-realisasjoner av I_n og deretter beregne den empiriske varians for disse.

Skjematisk kan en sette opp dette slik, idet $\mathbf{X}^{(1)},...,\mathbf{X}^{(k)}$ er de simulerte realisasjoner av vektoren $\mathbf{X}$.

$$\begin{array}{llll} \mathbf{X}^{(1)} = (\mathbf{X}_1^{(1)},...,\mathbf{X}_n^{(1)}) & \text{gir} & I_n^{(1)} \equiv T_n(\mathbf{X}^{(1)}) \\ \mathbf{X}^{(2)} = (\mathbf{X}_1^{(2)},...,\mathbf{X}_n^{(2)}) & \text{gir} & I_n^{(2)} \equiv T_n(\mathbf{X}^{(2)}) \\ \quad\vdots & \vdots & \vdots \\ \mathbf{X}^{(k)} = (\mathbf{X}_1^{(k)},...,\mathbf{X}_n^{(k)}) & \text{gir} & I_n^{(k)} \equiv T_n(\mathbf{X}^{(k)}) \end{array}$$

La nå

$$I_n^{(\cdot)} = \frac{1}{k} \sum_{l=1}^{k} I_n^{(l)}$$

Anslaget for $Var_{\theta_0} I_n$ er da gitt ved den empiriske varians for de simulerte verdier $I_n^{(l)}$, nemlig

$$(5.1) \quad \frac{1}{k-1} \sum_{l=1}^{k} (I_n^{(l)} - I_n^{(\cdot)})^2$$

Når $k \to \infty$ vil dette konvergere mot den eksakte verdi for $Var_{\theta_0} I_n$. Dette følger av at empirisk varians er en konsistent estimator for varians. (Vi forutsetter her at variansen virkelig eksisterer, og at vår random generator virker tilfredsstillende). Siden vi selv kan velge k, kan vi estimere $Var_{\theta_0} I_n$ med så stor nøyaktighet vi ønsker. Merk at n hele tida holdes fast, lik den aktuelle utvalgsstørrelsen.

I de to spesielle tilfellene vi ser på her, prisindeks for bruktbolig og prisindeks for nye eneboliger, må det simuleres realisasjoner av observasjoner både fra det aktuelle kvartalet og observasjoner fra de to foregående år. Dette fordi basis er beregnet på grunnlag av observasjoner fra foregående år og regresjonskoeffisientene beregnet på grunnlag av observasjoner fra de to foregående år.

For å få riktig anslag på variansen må variansen på grunn av frafallet i undersøkelsen beregnet ved $\hat{Var}_{\theta_0} I_n$ multipliseres med en faktor $\frac{N-n}{N-1}$.

I prisindeks for nye eneboliger er N lik alle fullførte eneboliger i kvartalet, mens n er alle eneboliger som er benyttet i indeksberegningen. I prisindeks for bruktbolig er N lik alle omsatte brukte boliger i kvartalet, mens n er alle omsatte brukte boliger som benyttes i indeksberegningen. Indeksens standardavvik er dermed gitt som $\sqrt{\hat{Var}_{\theta_0} I_n \cdot \frac{N-n}{N-1}}$

5.1. Bootstrapping

Metoden er blant annet beskrevet i Efron and Tibshirani (1993) og Lindqvist (1996).

Bootstrap - prinsippet går ut på at vi gjennomgår de samme rutinene som ovenfor, men med den ovenfor kjente sannsynlighetsfordelingen **F** erstattet med en som er estimert ut fra observasjonene $\mathbf{x} = (\mathbf{x}_1,...,\mathbf{x}_n)$. Det er ulike måter å gjøre dette på. I hovedsak skilles det mellom *parametrisk og ikke-parametrisk* bootstrap.

5.1.1. Parametrisk bootstrap

Her estimerer man først θ ut fra de opprinnelige observasjonene $\mathbf{x} = (\mathbf{x}_1,...,\mathbf{x}_n)$. La oss kalle estimatoren $\hat{\theta}$. Denne kan f.eks være maximum likelihood estimatoren (MLE). Deretter går vi fram nøyaktig som i tilfellet med kjent θ_0, bare at θ_0 er erstattet med estimatet $\hat{\theta}$.

Siden vi nå ikke trekker fra den virkelige underliggende fordelingen **F**, men fra den estimerte modellen $\hat{F}$ er det vanlig å sette en * på de genererte observasjonene av **X**, I_n, etc. Vi får følgende skjema, der altså $\mathbf{X}^{*(1)},...,\mathbf{X}^{*(k)}$ er de simulerte realisasjoner av vektoren $\mathbf{X}^* = (\mathbf{X}_1^*,...,\mathbf{X}_n^*)$ med komponenter trukket fra $\hat{F}$).

$$\mathbf{X}^{*(1)} = (\mathbf{X}_1^{*(1)},...,\mathbf{X}_n^{*(1)}) \text{ gir } \quad I_n^{*(1)} \equiv I_n(\mathbf{X}^{*(1)})$$
$$\mathbf{X}^{*(2)} = (\mathbf{X}_1^{*(2)},...,\mathbf{X}_n^{*(2)}) \text{ gir } \quad I_n^{*(2)} \equiv I_n(\mathbf{X}^{*(2)})$$
$$\vdots$$
$$\mathbf{X}^{*(k)} = (\mathbf{X}_1^{*(k)},...,\mathbf{X}_n^{*(k)}) \text{ gir } \quad I_n^{*(k)} \equiv I_n(\mathbf{X}^{*(k)})$$

La nå

$$I_n^{*(\cdot)} = \frac{1}{k}\sum_{l=1}^{k} I_n^{*(l)}$$

Vi definerer da

$$(5.2)\quad (\mathrm{Var}_\theta I_n)_{BOOT} = \frac{1}{k-1}\sum_{l=1}^{k} (I_n^{*(l)} - I_n^{*(\cdot)})^2$$

Praksis har vist at antallet *k* av Bootstrap-utvalg $\mathbf{X}^{*(l)}$ ikke behøver å være så stort for å få et brukbart variansestimat. Vanlig brukte verdier er fra 50 til 200, men helt ned i 20 kan fungere bra.

Merk at selv om $k \to \infty$ i (5.2)), vil høyresiden ikke konvergere mot $Var_{\theta_0} I_n$ (der θ_0 er den sanne verdi på parameteren), men mot estimatet $Var_{\hat{\theta}} I_n$, som blir en funksjon av de opprinnelige observasjonene $x_1,...,x_n$. Dette kommer selvsagt av at «observasjonene» i $\mathbf{X}^*$ er trukket fra $\hat{F}$). Merk også at *n* hele tida er fast, lik utvalgsstørrelsen i vårt opprinnelige utvalg.

Estimatet i (5.2) vil pga. bootstrap-trekningen variere dersom vi går gjennom det samme programmet flere ganger, selv om *k* ikke endres. For å få en entydig definisjon av bootstrap-estimatoren for varians er dermed

$$(5.3)\quad (\mathrm{Var}_\theta I_n)_{BOOT} = Var_{\hat{\theta}}\, I_n$$

5.1.2. Vanlig (ikke-parametrisk) bootstrap

Her genereres de nye datasettene $\mathbf{X}^{*(l)} = (\mathbf{X}_1^{*(l)},..., \mathbf{X}_n^{*(l)})$ (l = 1,...,k) fra den såkalte empiriske fordeling basert på de opprinnelige observasjonene $\mathbf{x}_1,..., \mathbf{x}_n$.

Den empiriske fordelingen er definert som en diskret fordeling med mulige verdier gitt ved de opprinnelige observasjonene $\mathbf{x}_1,..., \mathbf{x}_n$ og med sannsynlighet 1/n for hver. Det er igjen vanlig å bruke $\mathbf{X}^*$ som navn på en stokastisk variabel ved denne fordelingen. Skjematisk kan vi skrive denne fordelingen som

Mulige verdier for $\mathbf{X}^*$	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_n$
P($\mathbf{X}^*$ =)	1/n	1/n	...	1/n

det vil si at vi har en *uniform* sannsynlighetsmodell med utfallsrom $\{\mathbf{x}_1,...,\mathbf{x}_n\}$.

I og med at $\mathbf{X}^*$ har en uniform sannsynlighetsfordeling over $\{\mathbf{x}_1,...,\mathbf{x}_n\}$, kan uavhengige realisasjoner av denne simuleres ved tilfeldig trekking *med* tilbakelegging fra mengden $\{\mathbf{x}_1,...,\mathbf{x}_n\}$. En vektor $\mathbf{X}^{*(l)}$ simuleres derfor ved å trekke *n* ganger med tilbakelegging fra denne mengden. Når dette er gjort, og vi har beregnet k verdier $I_n^{*(1)},...,I_n^{*(k)}$, gir formel (5.2) variansestimatet $(\mathrm{Var}I_n)_{BOOT}$.

Som for den parametriske bootstrap vil den presise definisjonen av bootstrap-estimatoren for varians være grensen for (5.2) når $k \to \infty$. Siden $\mathbf{X}^*$ene nå trekkes fra den empiriske fordelingen for $\mathbf{x}_1,...,\mathbf{x}_n$ istedenfor fra F, vil dette endre (5.3) til

$$(5.4)\quad (\mathrm{Var}_\theta I_n)_{BOOT} = \mathrm{Var}_*(I_n^*)$$

der * som indeks på Var betegner at variansen er beregnet med hensyn på bootstrapfordelingen (her den empiriske fordelingen over $\{\mathbf{x}_1,...,\mathbf{x}_n\}$) og $I_n^* = I_n(\mathbf{X}^*)$.

5.2. Jackknife

Metoden er blant annet beskrevet i Efron and Tibshirani (1993).

Jackknife ligner mye på bootstrapping. Måten en genererer de nye datasettene på er noe annerledes. La $\hat{\theta}$ være en estimator for θ basert på alle observasjonene $\mathbf{X}_1, \mathbf{X}_2,..., \mathbf{X}_n$. Ved jackknife utelates en og en observasjon etter tur:

$$\mathbf{x}_{(i)} = (\mathbf{x}_1, \mathbf{x}_2,...\mathbf{x}_{i-1}, \mathbf{x}_{i+1},...\mathbf{x}_n)$$

for i = 1,2,...n. I'te jackknife-utvalg består av alle data unntatt observasjon nr. i. La

$$\hat{\theta}_{(i)} = I(\mathbf{x}_{(i)})$$

være i'te realisasjon av $\hat{\theta}$. Jackknife-estimatet for skjevhet er definert som

$$bias_{jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$$

hvor

$$\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)} / n.$$

Jackknife-estimatet for standardavvik er definert som

$$\hat{\sigma}_{jack} = \left[\frac{n-1}{n}\sum(\hat{\theta}_{(i)} - \theta_{(\cdot)})^2\right]^{1/2}$$

5.3. Kryssvalidering

Denne metoden er blitt brukt for å gi et anslag for standardavviket til boligprisindeksene og er gitt i Lillegård (1994). Anta at vi deler datamaterialet, kvartalet og basis, i for eksempel fire like store grupper eller utvalg. (Vi kunne valgt et annet tall, men anta nå at vi deler inn i fire). Inndelingen skjer tilfeldig. Deretter beregner vi fire prisindekser, en for hver av de fire gruppene. La oss kalle prisindeksene for Y_1, Y_2, Y_3 og Y_4. Den totale prisindeksen I, den vi ville fått ved å bruke hele datamaterialet, er omtrent lik gjennomsnittet av de fire enkeltindeksene. Prisindeksen kan altså beregnes ved

$$I = \tfrac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$$

Hvis vi antar at de fire enkeltindeksene er uavhengige, vil variansen til gjennomsnittet bli en fjerdedel av variansen til de fire enkeltindeksene. Hvis enkeltindeksene alle har varians lik σ_Y^2, blir σ_I^2, variansen til totalindeksen, lik

$$\sigma_I^2 = \tfrac{1}{16}[Var(Y_1) + Var(Y_2) + Var(Y_3) + Var(Y_4)]$$
$$= \frac{4\sigma_Y^2}{16} = \frac{\sigma_Y^2}{4}$$

Dermed halveres standardavviket

$$\sigma_I = \sqrt{\sigma_I^2} = \frac{\sigma_Y}{2}$$

Ettersom vi har fire observasjoner kan vi estimere standardavviket til enkeltindeksene med kjente metoder. Hvis vi kaller de estimerte standardavvikene for S, får vi følgende formel for standardavviket til totalindeksen

$$S_I = \frac{S_Y}{2} = \frac{\sqrt{\frac{1}{3}\sum_{i=1}^{4}(Y_i - \overline{Y})^2}}{2}$$

Dersom vi ønsker å legge mer enn en periode (flere kvartaler) til grunn for beregning av variansen kan dette gjøres som følger. Anta at vi ønsker å legge *m* perioder til grunn og estimerer variansen til prisindeksen i periode nr. *j* ved

$$S_j^2 = \frac{\frac{1}{3}\sum_{i=1}^{4}(Y_{ij} - \overline{Y}_j)^2}{4}; \text{ j=1,2,...,m}$$

hvor Y_{ij} er observasjon nr. *i* av prisindeksen i kvartal nr. *j*.

Man antar så at en prisindeks har samme varians i alle kvartaler. Forskjeller som f.eks. skyldes ulikt antall observasjoner antas å være neglisjerbare. Variansen til prisindeksen blir derfor gjennomsnittet av de estimerte variansene i hver periode

$$S^2 = \tfrac{1}{m}\sum_{j=1}^{m}\sigma_j^2$$

5.3.1. Antall grupper

I Lillegård (1994) er det foreslått å dele datamaterialet inn i fire grupper. Det vil være av interesse å se om en inndeling av datamaterialet i et annet antall grupper vil gi samme estimat på standardavviket. Dersom vi generelt deler datamaterialet inn i *n* grupper får vi på tilsvarende måte at prisindeksen kan beregnes ved

$$I = \tfrac{1}{n}(Y_1 + Y_2 + ... + Y_n)$$

Variansen til gjennomsnittet blir en n'te-del av variansen til enkeltindeksene:

$$\sigma_I^2 = \tfrac{1}{n^2}[Var(Y_1) + Var(Y_2) + ... + Var(Y_n)] = \frac{n\sigma_Y^2}{n^2} = \frac{\sigma_Y^2}{n}$$

Videre får man følgende sammenheng for standardavviket

$$\sigma_I = \sqrt{\sigma_I^2} = \frac{\sigma_Y}{\sqrt{n}}$$

Spesielt gir dette

$$\sigma_I = \sqrt{\sigma_I^2} = \frac{\sigma_Y}{2} \quad (n = 4)$$

$$\sigma_I = \sqrt{\sigma_I^2} = \frac{\sigma_Y}{3} \quad (n = 9)$$

6. Intervallestimering

Dersom vi har et estimat for variansen, har vi også muligheten for å lage konfidensintervaller for den estimerte indeksverdien. Et intervallestimat er ofte mer nyttig enn et punktestimat. Vi kan da komme fram til et intervall som "med en viss sikkerhet" inneholder den ukjente indeksen. Det er her beskrevet noen av de mulighetene vi har for å lage konfidensintervaller.

6.1. Standardintervallet

Gitt et estimat for $\hat{\theta}$ og et estimat for variansen $\hat{\sigma}^2$, vil disse til sammen gi et konfidensintervall $\hat{\theta} \pm \hat{\sigma} z^{\alpha}$, hvor z^{α} er kvantilen i en standard normalfordeling. Intervallet vil inneholde den ukjente indeksen med sannsynlighet lik 1 -2α. Vi kaller dette standardintervallet for θ.

Ved bruk av asymptotisk teori vil man ha, at dersom utvalgsstørrelsen *n* blir stor, vil fordelingen for $\hat{\theta}$ blir mer og mer normal, med forventning nær θ og varians nær $\hat{\sigma}^2$.

6.2. Bootstrap konfidensintervall

La situasjonen være som i kapittel 5. Som i kapittel 5.1.2 lager vi bootstrap-simuleringer av $\hat{\theta}$, $\hat{\theta}^{*(1)}$,..., $\hat{\theta}^{*(k)}$, der *k* er et valgt (stort) tall. For konfidensintervaller bør dette være større enn for variansestimering, f.eks. k = 1 000 eller mer.

De simulerte verdier $\hat{\theta}^{*(l)}$ sorteres nå i stigende rekkefølge. Vi vil anta i det følgende at

$$\hat{\theta}^{*(1)} \leq \hat{\theta}^{*(2)} \leq \ldots \leq \hat{\theta}^{*(k)}$$

6.2.1. Persentilmetoden

Anta at vi ønsker et konfidensintervall for θ med konfidenskoeffisient 1 - α. *Persentilintervallet* er nå gitt ved

$$[\hat{\theta}^{*(L)}, \hat{\theta}^{*(U)}]$$

der

$$L = [\frac{\alpha}{2}k] \text{ og } U = [(1 - \frac{\alpha}{2})k]$$

([α betyr heltallsdelen av α, dvs. f.eks. [5,37] = 5).

Med ord betyr dette at vi kutter andelen α/2 av de laveste og α/2 av de høyeste verdiene for θ*, og lar konfidensintervallet være det intervallet av verdier som da gjenstår. Hvis vi ser på et histogram for $\theta^{*(l)}$, $\theta^{*(2)},\ldots,\theta^{*(k)}$, vil arealene på henholdsvis venstre side av $\hat{\theta}^{*(L)}$ og høyre side av $\hat{\theta}^{*(U)}$ begge være (ca.) α/2.

6.2.2. BC-intervallet

Dette er en forbedret metode for bootstrap-konfidensintervaller. (BC = «Bias Corrected»). Man estimerer først «bias-konstanten» z_0 fra relasjonen

$$(6.1) \quad \Phi(z_0) = P_*(\theta^* < \hat{\theta})$$

Her er Φ som vanlig fordelingsfunksjonen i N(0,1) mens * i P_* betegner som i formel (5.4) at sannsynligheten beregnes med hensyn på bootstrap-fordelingen (enten denne er parametrisk eller ikke-parametrisk). Vi ser at $\Phi(z_0) = 0{,}50$ svarer til at θ er medianen i bootstrap-fordelingen til θ*. Dersom sannsynlighetsfordelingen for $\hat{\theta}^*$ er skjev i forhold til $\hat{\theta}$ ($P_*(\hat{\theta}^* < \hat{\theta}) \neq 0{,}50$), tyder det på at den opprinnelige estimatoren θ har en usymmetrisk fordeling, og BC-intervallet skal justere for denne skjevheten.

Høyresiden i (6.1) kan estimeres ut fra histogrammet for $\hat{\theta}^*$ (arealet opp til $\hat{\theta}$), eller fra de ordnede $\hat{\theta}^{*(l)}$-verdier ved å telle opp antall $\hat{\theta}^{*(l)}$ som er $< \hat{\theta}$ og dele på *k*):

$$P_*(\hat{\theta}^* < \hat{\theta}) \approx \frac{l}{k}$$

der *l* er størst mulig med $\hat{\theta}^{*(l)} < \hat{\theta}$. Nå vil z_0 finnes fra (6.1) ved å bruke tabell eller dataprogram for Φ^{-1}.

BC-metoden modifiserer persentilmetoden ved at vi velger

$L_{BC} = [\Phi(2z_0 - u_{\alpha/2})k]$ og $U_{BC} = [\Phi(2z_0 + u_{\alpha/2})k]$

og lar konfidensintervallet være

$[\hat{\theta}^{*(L_{BC})}, \hat{\theta}^{*(U_{BC})}]$

Legg merke til at $z_0 = 0$ gir persentilintervallet.

7. Numeriske resultater

Metoden med bootstrapping gir estimat som konvergerer mot standardavviket for økende antall bootstrap-utvalg. Som en test på hvor mange bootstrap-utvalg vi trenger, har vi estimert standardavviket i prisindeks for nye eneboliger i 4. kvartal 1997 for 10, 20 ,50, 100, 200 og 500 bootstrap-utvalg. Resultatene er vist i tabell 7.1. Denne testen viser at det er tilstrekkelig med 100 - 200 bootstrap-utvalg for å få et tilfredsstillende estimat på standardavviket. Man kunne kanskje greid seg med 100 utvalg, men resultatet vil generelt bli bedre jo flere utvalg man trekker. Spesielt ser man at estimat basert på 10 og 20 utvalg blir mer usikre. Vi velger derfor å trekke 200 utvalg ved bruk av bootstrapping.

Metodene jackknife, bootstrap, inndeling av datamaterialet i fire og ni grupper er alle benyttet til å estimere standardavviket i prisindeks for nye eneboliger i 4. kvartal 1997. Resultatene er gitt i tabell 7.2.

Alle fire estimatene for standardavviket er relativt like. Metodene med inndeling av datamaterialet i fire og ni grupper virker imidlertid veldig ustabile. Ved å gjøre beregningene 20 ganger ble minimums- og maksimumsverdien for estimert standardavvik ved inndeling i fire grupper henholdsvis 0,2 og 1,2. Ved inndeling i ni grupper ble tilsvarende verdier 0,3 og 1,0. Det resultatet man får fra gang til gang blir noe tilfeldig. Et gjennomsnitt av disse 20 simuleringene (som er gitt i tabell 7.2) gir imidlertid et estimat som ligger i samme størrelsesorden som de en får ved jackknife og bootstrap. Det kan imidlertid være litt tilfeldig at resultatene stemmer så godt overens. Metodene jackknife og bootstrap synes være de to beste metodene å benytte. Disse to metodene er derfor benyttet til å estimere standardavviket til prisindeks for nye eneboliger i alle kvartaler i 1996 og 1997. Resultatene er gitt i tabell 7.3.

Metodene bootstrap og jackknife synes å gi like resultater i alle kvartaler. Bootstrap-metoden er imidlertid mer hensiktsmessig å bruke da en her kun trenger 200 bootstrap-utvalg, mens antall jackknife-utvalg alltid må være lik størrelsen på datasettet. I prisindeks for nye eneboliger ligger dette i størrelsesorden 800 - 1 300 observasjoner per kvartal (se tabell 7.4). Bootstrap-metoden blir mindre beregningskrevende, og er derfor å foretrekke.

Tabell 7.1. Estimert standardavvik for estimerte indeksverdier i prisindeks for nye eneboliger i 4. kvartal 1997 ved bruk av bootstrapping

Antall utvalg	Estimert standardavvik
10 utvalg	0,8
20 utvalg	0,6
50 utvalg	0,6
100 utvalg	0,7
200 utvalg	0,7
500 utvalg	0,7

Tabell 7.2. Estimert standardavvik for estimerte indeksverdier ved bruk av bootstrapping, jackknife, inndeling av datasettet i fire grupper og inndeling av datasettet i ni grupper. Prisindeks for nye eneboliger i 4. kvartal 1997

Kvartal	Jackknife	Bootstrap	Inndeling i 4 grupper	Inndeling i 9 grupper
4. kv. 1997	0,7	0,7	0,7	0,6

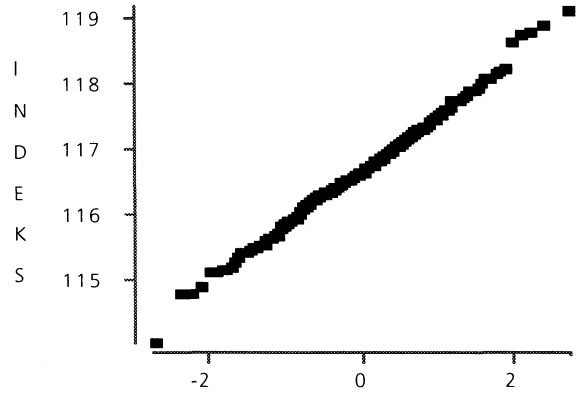
Tabell 7.3. Estimert standardavvik for estimerte indeksverdier ved bruk av bootstrap og jackknife. Prisindeks for nye eneboliger i 1996 og 1997

Kvartal	Jackknife	Bootstrap
1. kv. 1996	0,5	0,5
2. kv. 1996	0,6	0,6
3. kv. 1996	0,7	0,7
4. kv. 1996	0,5	0,5
1. kv. 1997	0,7	0,7
2. kv. 1997	0,7	0,7
3. kv. 1997	0,7	0,7
4. kv. 1997	0,7	0,7

Tabell 7.4. Antall observasjoner i populasjonen og antall observasjoner brukt i indeksberegningene. Prisindeks for nye eneboliger i 1996 og 1997

Kvartal	Antall obs. brukt i indeksberegning	Antall obs. i populasjonen	Prosentvis andel av populasjonen brukt i indeksberegningen
1. kv. 1996:	941	1 752	54
2. kv. 1996:	906	1 816	50
3. kv. 1996:	836	1 752	48
4. kv. 1996:	1 269	2 468	51
1. kv. 1997:	815	1 738	47
2. kv. 1997:	862	1 814	48
3. kv. 1997:	769	1 659	46
4. kv. 1997:	1 231	2 666	46

Figur 7.1. Normalplott av estimerte indeksverdier i prisindeks for nye eneboliger i 4. kvartal 1997



Resultatene viser videre at det estimerte standardavviket ligger mellom 0,5 og 0,7 i de åtte kvartalene i 1996 og 1997. Vi ser av tabell 7.4 at standardavviket er minst i de kvartalene hvor den prosentvise andelen av populasjonen brukt i indeksberegningen er størst, det vil si at den estimerte indeksverdien er mer sikker, jo mindre frafall vi har i undersøkelsen. Også antall observasjoner som ligger til grunn for indeksberegningene er av betydning for indeksenes standardavvik. Jo flere observasjoner som ligger til grunn, jo mindre usikker blir den estimerte indeksverdien.

Som vist i kapittel 2 kan man, dersom man har estimert et standardavvik std($\hat{I}^{k,T}$) for estimatoren $\hat{I}^{k,T}$, finne et konfidensintervall for estimatoren. I tabell 7.5 er det beregnet et 90 prosent konfidensintervall for estimerte indeksverdier i prisindeks for nye eneboliger i 1996 og 1997.

Konfidensintervallene i tabell 7.5 er beregnet ut fra antakelsen om at fordelingene til indeksestimatorene er normalfordelte. Ved å lage et normalplott av simulerte indeksverdier kan vi se om fordelingen til

Tabell 7.5. 90 prosent konfidensintervall for estimerte indeksverdier i prisindeks for nye eneboliger i 1996 og 1997. Standardavvik estimert ved bootstrapping

Kvartal	Konfidensintervall
1. kv. 1996:	108.1±0.8 = [107.3, 108.9]
2. kv. 1996:	109.1±1.0 = [108.1, 110,1]
3. kv. 1996:	108.7±1.1 = [107.5, 109.1]
4. kv. 1996:	108.3±0.8 = [107.5, 109.1]
1. kv. 1997:	109.4±1.1 = [108.3, 110.5]
2. kv. 1997:	112.6±1.1 = [111.5, 113.7]
3. kv. 1997:	114.9±1.1= [113.8, 116.0]
4. kv. 1997:	116.8±1.1 = [115.7, 117.9]

Tabell 7.6. Estimert standardavvik for endringsestimatoren for to etterfølgende kvartaler i prisindeks for nye eneboliger i 1996 og 1997. Standardavvik estimert ved bootstrapping

Kvartal	Standardavvik
2. kvartal 1996-1. kvartal 1996	0,8
3. kvartal 1996-2. kvartal 1996	0,9
4. kvartal 1996-3. kvartal 1996	0,9
1. kvartal 1997-4. kvartal 1996	0,9
2. kvartal 1997-1. kvartal 1997	1,0
3. kvartal 1997-2. kvartal 1997	1,0
4. kvartal 1997-3. kvartal 1997	1,0

indeksestimatoren er normalfordelt. Observasjonene bør da ligge på ei tilnærmet rett linje som ligger i 45 graders vinkel i forhold til x-aksen. Som en test på dette viser figur 7.1 et normalplott av simulerte indeksverdier fra 4. kvartal 1997. De simulerte indeksverdiene synes å være tilnærmet normalfordelt. Et par tester bekrefter også at intervallestimering ved bruk av bootstrap-konfindensintervaller, som er beskrevet i kapittel 6.2, gir tilnærmet samme resultat som ved antakelse om normalfordeling. I og med at beregning av bootstrap-konfidensintervaller krever minst 1 000 bootstrap-simuleringer for å gi et godt resultat, velger vi å beregne konfidensintervallene ved bruk av antakelsen om normalfordeling. Vi greier oss da med 200 simuleringer.

Resultatene i tabell 7.5 gir oss for eksempel at vi er 90 prosent sikre på at den virkelig indeksverdien for 1. kvartal 1996 er mellom 107,3 og 108,9. Dette forutsatt at vi har ingen andre feilkilder enn frafall.

I tillegg til å se på standardavviket til estimatorene for enkeltindekser, er det også interessant å se på variasjonen i endringsestimatene mellom to etterfølgende kvartaler. For å estimere standardavviket til endringsestimatoren $\hat{\Delta}^{k,T}$ antar vi følgende sammenheng:

$$std(\hat{\Delta}^{k,T}) = \sqrt{Var(\hat{\Delta}^{k,T})} = \sqrt{Var(\hat{I}^{k+1,T} - I^{k,T})}$$
$$= \sqrt{Var(\hat{I}^{k+1,T}) + Var(\hat{I}^{k,T})}$$

Utregningene i tabell 7.6 er gjort for endringer i indeksverdier i prisindeks for nye eneboliger i 1996 og 1997.

Tabell 7.7 viser 90 prosent konfidensintervaller for endringsestimatoren i to etterfølgende kvartaler i 1996 og 1997. Tabellen viser for eksempel at vi er 90 prosent sikre på at den ukjente endringen i indeksverdier mellom 1. og 2. kvartal 1996 ligger mellom -0,3 og 2,3. Siden dette konfidensintervallet inneholder verdien 0, konkluderer vi med at endringen ikke er signifikant på 10 prosent nivå. Endringen fra 1. til 2. kvartal 1997 er imidlertid signifikant på 10 prosent nivå, da dette intervallet ikke dekker 0. Dette forutsetter som nevnt tidligere antakelsen om at det ikke er andre feilkilder enn frafall.

Det er også gjort tilsvarende analyser for prisindeks for bruktbolig. Standardavviket er beregnet ved hjelp av bootstrapping, med 200 bootstrap-utvalg. Estimert standardavvik for estimerte indeksverdier er gitt i tabell 7.8. Tabellen viser at standardavviket til totalindeksen ligger mellom 0,3 og 0,4 prosentpoeng. Standardavviket til totalindeksene for eneboliger, småhus og blokkleiligheter ligger alle mellom 0,5 og 0,7 prosentpoeng. Ingen av totalindeksene for de ulike prissonene ligger over 1,5 prosentpoeng. Noen av delindeksene er imidlertid mer usikre.

I figur 7.2 til 7.5 er det estimerte standardavviket for de fire kvartalene i 1997 plottet for de ulike delindeksene. Figurene viser at standardavviket er forholdsvis stabilt for alle delindeksene over tid. En ser også at standardavviket gjennomgående er størst i de prissonene hvor det er færrest prisobservasjoner. For eneboliger er standardavviket størst i prissonen Oslo m/Bærum, hvor standardavviket ligger mellom 2,4 og 3,2 prosentpoeng. Her ligger antall prisobservasjoner hvert kvartal mellom 150 og 250 (se tabell 7.10). For eneboliger i prissonen "resten av landet" , hvor antall prisobservasjoner er godt over 1 000 i alle kvartaler, ligger standardavviket på rundt 0,5 prosentpoeng. For blokkleiligheter derimot, er det Oslo m/Bærum som har flest prisobservasjoner. Her ligger antall prisobservasjoner hvert kvartal i størrelsesorden 500 – 1 000, og standardavviket ligger mellom 0,5 og 1,1 prosentpoeng. Prissonen "resten av Akershus" har færrest prisobservasjoner for denne boligtypen, med et standardavvik mellom 1,5 og 2,0 prosentpoeng.

I de tilfellene hvor delindeksene har høyt standardavvik blir sammenligningen fra kvartal til kvartal mer usikker. Den langsiktige trenden er alltid mer pålitelig.

I figur 7.6 er estimert standardavvik for de ulike delindeksene fra alle fire kvartaler i 1997 plottet mot antall observasjoner brukt i indeksberegningene. Figuren bekrefter at standardavviket synker med økende antall prisobservasjoner. Det ser for eksempel ut som man må

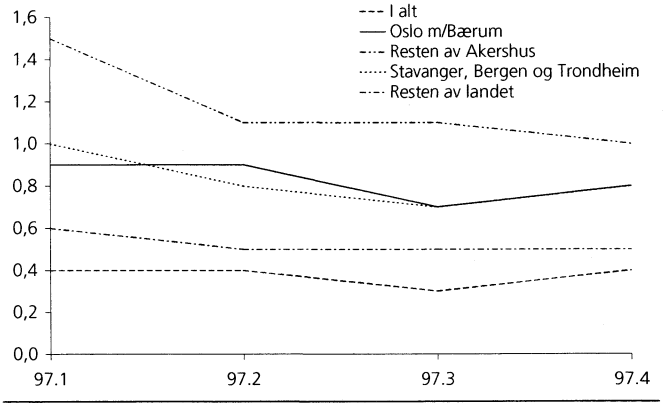
Tabell 7.7. Estimerte 90 prosent konfidensintervaller for endringsestimatoren i to etterfølgende kvartaler i prisindeks for nye eneboliger i 1996 og 1997. Standardavvik estimert ved bootstrap

Kvartal	Konfidensintervall
2. kvartal 1996 - 1. kvartal 1996	1.0±1.3 = [-0.3, 2.3]
3. kvartal 1996 - 2. kvartal 1996	-0.4±1.4 = [-1.8, 1.0]
4. kvartal 1996 - 3. kvartal 1996	-0.4±1.4 = [-0.3, 2.5]
1. kvartal 1997 - 4. kvartal 1996	1.1±1.4 = [-0.3, 2.5]
2. kvartal 1997 - 1. kvartal 1997	3.2±1.6 = [1.6, 4.8]
3. kvartal 1997 - 2. kvartal 1997	2.3±1.6 = [0.7, 3.9]
4. kvartal 1997 - 3. kvartal 1997	1.9±1.6 = [0.3, 3.5]

Tabell 7.8. Estimert standardavvik for estimerte indeksverdier ved bruk av bootstrap. Prisindeks for bruktbolig i 1997

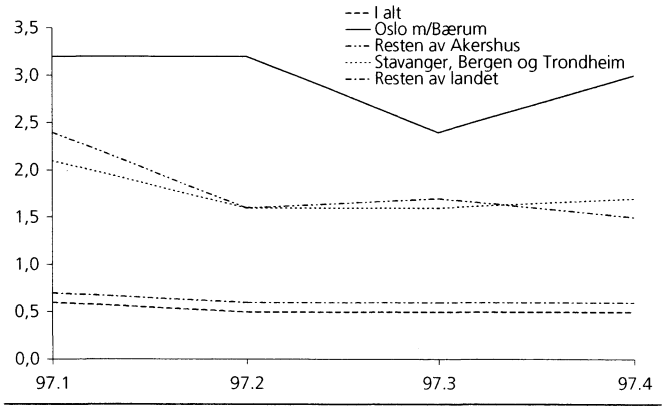
	I alt	Oslo m/ Bærum	Resten av Akershus	Stavanger, Bergen og Trondheim	Resten av landet
Bruktbolig i alt					
1. kvartal	0,4	0,9	1,5	1,0	0,6
2. kvartal	0,4	0,9	1,1	0,8	0,5
3. kvartal	0,3	0,7	1,1	0,7	0,5
4. kvartal	0,4	0,8	1,0	0,8	0,5
Eneboliger					
1. kvartal	0,6	3,2	2,4	2,1	0,7
2. kvartal	0,5	3,2	1,6	1,6	0,6
3. kvartal	0,5	2,4	1,7	1,6	0,6
4. kvartal	0,5	3,0	1,5	1,7	0,6
Småhus					
1. kvartal	0,7	1,8	1,7	1,3	0,9
2. kvartal	0,6	2,1	1,5	0,9	0,8
3. kvartal	0,5	1,6	1,4	1,0	0,7
4. kvartal	0,6	1,7	1,6	1,1	0,8
Blokkleiligheter					
1. kvartal	0,7	1,1	1,9	1,8	1,3
2. kvartal	0,6	0,9	2,0	1,4	1,1
3. kvartal	0,5	0,7	1,5	1,2	0,9
4. kvartal	0,6	0,8	1,8	1,4	1,0

Figur 7.2. Estimert standardavvik. Bruktbolig i alt

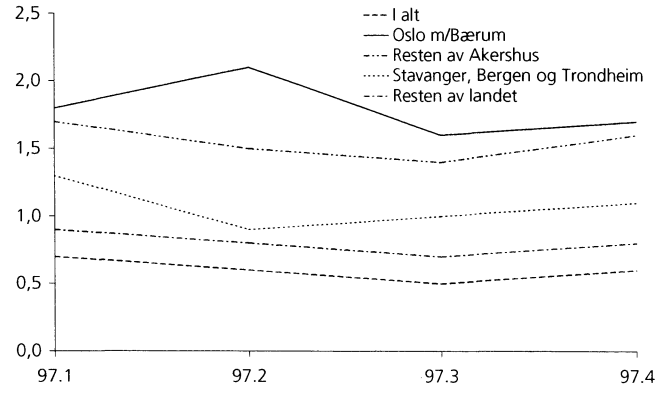


ha over 250 prisobservasjoner for at standardavviket ikke skal overstige 2 prosentpoeng. Figuren viser også at standardavviket ikke blir merkbart mindre når antall prisobservasjoner overstiger 1 000.

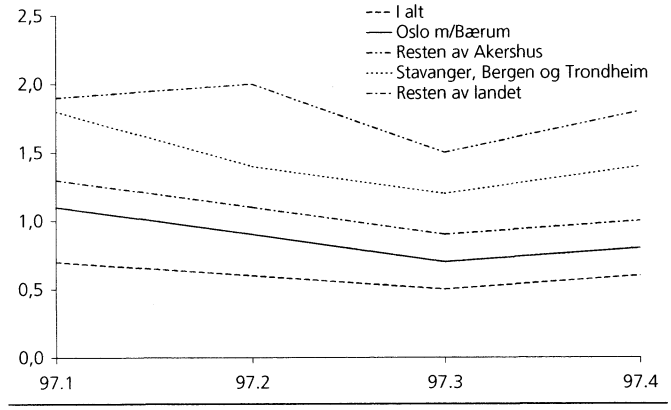
Figur 7.3. Estimert standardavvik. Enebolig



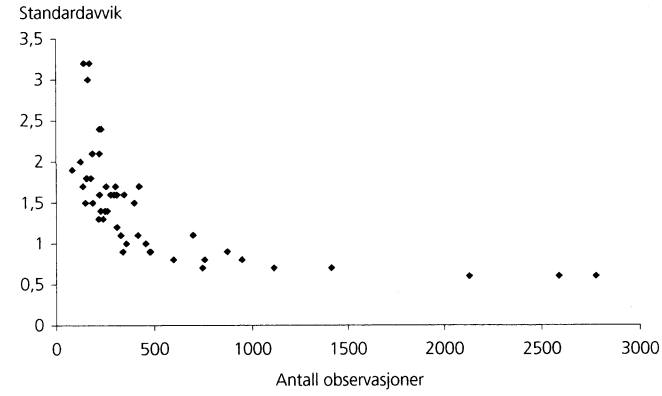
Figur 7.4. Estimert standardavvik. Småhus



Figur 7.5. Estimert standardavvik. Blokkleilighet



Figur 7.6. Estimert standardavvik. Alle boligtyper



Tabell 7.9. 90 % konfidensintervall for estimerte indeksverdier i prisindeks for bruktbolig i 1997. Standardavvik estimert ved bootstrapping

	I alt	Oslo m/ Bærum	Resten av Akershus	Stavanger, Bergen og Trondheim	Resten av landet
Bruktbolig i alt					
1. kvartal	[131.6, 132.8]	[151.7, 154.5]	[133.4, 138.2]	[139.1, 142.3]	[120.1, 122.1]
2. kvartal	[136.2, 137.4]	[161.2, 164.0]	[142.0, 145.6]	[144.1, 146.7]	[123.3, 124.9]
3. kvartal	[139.2, 140.2]	[165.4, 167.6]	[146.3, 149.9]	[147.1, 149.3]	[125.9, 127.5]
4. kvartal	[140.9, 142.1]	[169.2, 171.8]	[151.0, 154.2]	[147.7, 150.3]	[127.0, 128.6]
Eneboliger					
1. kvartal	[126.4, 128.4]	[158.0, 168.2]	[133.4, 141.0]	[141.6, 148.4]	[119.8, 122.0]
2. kvartal	[131.1, 132.7]	[168.5, 178.7]	[146.7, 151.9]	[146.1, 151.3]	[123.5, 125.5]
3. kvartal	[134.3, 135.9]	[181.9, 189.5]	[151.0, 156.4]	[146.5, 151.7]	[126.4, 128.4]
4. kvartal	[135.8, 137.4]	[184.3, 192.9]	[157.5, 162.3]	[152.3, 157.7]	[127.1, 129.1]
Småhus					
1. kvartal	[132.7, 134.9]	[159.7, 165.5]	[132.7, 138.1]	[135.5, 139.7]	[121.8, 124.6]
2. kvartal	[134.8, 136.8]	[168.3, 175.1]	[134.0, 138.8]	[141.1, 143.9]	[121.3, 123.9]
3. kvartal	[138.8, 140.4]	[168.1, 173.3]	[140.5, 144.9]	[148.0, 150.6]	[125.2, 127.4]
4. kvartal	[138.4, 140.4]	[169.6, 175.0]	[141.0, 146.2]	[146.3, 148.5]	[124.6, 127.2]
Blokkleiligheter					
1. kvartal	[138.1, 140.3]	[146.6, 150.2]	[130.7, 136,7]	[138.1, 143.9]	[116.5, 120.7]
2. kvartal	[145.9, 147.9]	[156.4, 159.2]	[134.9, 141.3]	[144.5, 148.9]	[123.0, 126.2]
3. kvartal	[147.0, 148.6]	[160.6, 162.8]	[134.8, 139.6]	[144.9, 148.7]	[120.1, 122.9]
4. kvartal	[151.8, 153.8]	[165.7, 168.3]	[138.9, 144.7]	[143.7, 148.1]	[128.0, 131.2]

Tabell 7.10. Antall observasjoner brukt i indeksberegningene. Prisindeks for bruktbolig i 1997

	Oslo m/ Bærum	Resten av Akershus	Stavanger, Bergen og Trondheim	Resten av landet
Eneboliger				
1. kvartal	145	225	189	1 413
2. kvartal	174	348	297	2 588
3. kvartal	234	424	280	2 775
4. kvartal	166	400	304	2 126
Småhus				
1. kvartal	159	139	220	481
2. kvartal	225	189	342	756
3. kvartal	311	263	359	745
4. kvartal	257	224	331	597
Blokkleiligheter				
1. kvartal	697	84	180	242
2. kvartal	872	128	230	418
3. kvartal	1 113	152	312	477
4. kvartal	946	157	251	457

Tabell 7.11. Estimert standardavvik for endringsestimatoren for to etterfølgende kvartaler i prisindeks for bruktbolig i 1997. Standardavvik estimert ved bootstrapping

	I alt	Oslo m/ Bærum	Resten av Akershus	Stavanger, Bergen og Trondheim	Resten av landet
Bruktbolig i alt					
2. kvartal - 1. kvartal	0,6	1,3	1,9	1,3	0,8
3. kvartal - 2. kvartal	0,5	1,1	1,6	1,1	0,7
4. kvartal - 3. kvartal	0,5	1,1	1,5	1,1	0,7
Eneboliger					
2. kvartal - 1. kvartal	0,8	4,5	2,9	2,6	0,9
3. kvartal - 2. kvartal	0,7	4,0	2,3	2,3	0,8
4. kvartal - 3. kvartal	0,7	3,8	2,3	2,3	0,8
Småhus					
2. kvartal - 1. kvartal	0,9	2,8	2,3	1,6	1,2
3. kvartal - 2. kvartal	0,8	2,6	2,1	1,2	1,1
4. kvartal - 3. kvartal	0,8	2,3	2,1	1,1	1,1
Blokkleiligheter					
2. kvartal - 1. kvartal	0,9	1,4	2,8	2,3	1,7
3. kvartal - 2. kvartal	0,8	1,1	2,5	1,8	1,4
4. kvartal - 3. kvartal	0,8	1,1	2,3	1,8	1,3

Det er også her beregnet 90 prosent konfidensintervall for estimerte indeksverdier. Disse er gitt i tabell 7.9. Tabellen viser for eksempel at vi er 90 prosent sikre på at totalindeksen i 1. kvartal 1997 ligger mellom 131,6 og 132,8. Også her forutsetter vi at frafall er den eneste feilkilden.

Tabell 7.11 viser estimert standardavvik for endringsestimatoren for to etterfølgende kvartaler i 1997. På samme måte som tidligere kan man på grunnlag av det estimerte standardavviket finne et konfidensintervall for endringsestimatoren, og dermed se om den ukjente endringen i indeksverdi mellom to etterfølgende kvartaler er signifikant. For eksempel steg publisert totalindeks fra 1. til 2. kvartal 1997 med 4,6 prosentpoeng. Med et estimert standardavvik for denne endringsestimatoren på 0,6 prosentpoeng kan vi, dersom vi antar at frafall er den eneste feilkilden, være 90 prosent sikre på at den ukjente endringen mellom 1. og 2. kvartal 1997 ligger mellom 3,6 og 5,6 prosentpoeng. I og med at intervallet ikke dekker 0 betyr dette at endringen mellom 1. og 2. kvartal 1997 er signifikant på 10 prosent nivå.

8. Sammenligning av modellene

Når det skal vurderes hvilke av metodene som er å foretrekke for estimering av standardavviket er det flere forhold som bør legges til grunn. En bør blant annet vurdere metodens realisme, hvor raskt beregningene går og hvor plasskrevende de er.

Metodene jackknife og bootstrap er nokså like metoder. Ved bruk av jackknife må vi generelt beregne indeksen I for n jackknife-datasett, der n er antallet observasjoner som ligger til grunn for beregning av indeksen. Jackknife vil derfor være lettere å beregne hvis n er mindre enn de 200 bootstrap-datasettene som kreves for å estimere variansen ved bootstrapping. Med lettere å beregne mener vi at beregningene går raskere og at de er mindre plasskrevende.

Både i prisindeks for nye eneboliger og prisindeks for bruktboliger ligger det i de aller fleste tilfeller mer enn 200 observasjoner til grunn for beregning av indeksene. For boligprisindeksene vil derfor bootstrapping være å foretrekke framfor jackknife.

Ved bruk av metoden fra Lillegård (1994) ser en at estimatet for variansen blir veldig forskjellig fra gang til gang, dersom vi gjør beregningene flere ganger. En løsning på dette er å bruke et gjennomsnitt av et visst antall beregninger som estimat. Metoden synes imidlertid å være så ustabil, slik at bootstrapping vil være å foretrekke også framfor denne metoden.

9. Oppsummering og konklusjoner

Av de tre metodene som er benyttet til å estimere standardavviket til prisindeks for nye eneboliger og prisindeks for bruktbolig synes bootstrapping å være den beste metoden. Estimatene for standardavviket til prisindeks for nye eneboliger og totalindeksene i prisindeks for bruktbolig synes alle å være tilfredsstillende. Det er imidlertid vanskelig å gi noen eksakt grense for hva som er et tilfredsstillende standardavvik. Noen av delindeksene i prisindeks for bruktbolig har et noe høyt standardavvik. Dette gjelder i første rekke eneboliger i Oslo m/Bærum. Dette skyldes i stor grad at det er få prisobservasjoner som ligger til grunn for indeksberegningene, samt stor prisvariasjon mellom observasjonene. I slike tilfeller vil sammenligningen mellom to etterfølgende kvartaler være usikker. Den langsiktige trenden er alltid mer pålitelig.

For å opprettholde kvaliteten på indeksene er det viktig at svarprosenten på de utsendte spørreskjemaene ikke går ned i forhold til dagens nivå. Dette vil spesielt gjelde i perioder da det bygges få boliger eller omsettes få brukte boliger.

For framtida kan man samtidig med kvartalsvis frigiving av disse to statistikkene også publisere et estimert standardavvik for prisindeks for nye eneboliger og prisindeks for bruktbolig.

SAS-programmene som er benyttet til variansestimeringen inneholder noen generelle rutiner som kan være til nytte i andre sammenhenger. Programmene er derfor lagt ut på fellesdisken under området q:\dok\bolig\program.

Referanser

Efron, B. and R . Tibshirani (1993): *An Introduction to the Bootstrap*. Chapman & Hall.

Garthwaite, P. H., I. T Jolliffe and B. Jones (1995): *Statistical Inference*. Prentice Hall.

Lillegård, M. (1994): *Prisindekser for boligmarkedet*, Rapporter 94/7, Statistisk sentralbyrå.

Lindqvist, B. (1996): Bootstrapping, Forelesningsnotat i faget "EDB-intensiv statistikk", NTNU.

Wass, K. Å. (1992): *Prisindeks for ny enebolig*, Rapporter 92/21, Statistisk sentralbyrå.

De sist utgitte publikasjonene i serien Rapporter

Recent publications in the series Reports

Merverdiavgift på 23 prosent kommer i tillegg til prisene i denne oversikten hvis ikke annet er oppgitt

98/5 A.S. Bye og K. Mork: Resultatkontroll jordbruk 1998: Gjennomføring av tiltak mot forurensninger. 1998. 89s. 95 kr inkl. mva. ISBN 82-537-4397-1

98/6 K.R. Gerdrup: Skattesystem og skattestatistikk i et historisk perspektiv. 1998. 59s. 115 kr inkl. mva. ISBN 82-537-4531-1

98/7 E. Lofthus og Å. Osmunddalen: Innvandrere og sosialhjelp: Får mer fordi de trenger mer?. 1998. 32s. 100 kr inkl. mva. ISBN 82-537-4533-8

98/8 A. Langørgen og R. Aaberge: Gruppering av kommuner etter folkemengde og økonomiske rammebetingelser. 1998. 60s. 115 kr inkl. mva. ISBN 82-537-4535-4

98/9 A. Thomassen og R. Jensen: Kvadratmeterpriser for skolebygg. 1998. 24s. 100 kr inkl. mva. ISBN 82-537-4539-7

98/10 K. Ibenholt og H. Wiig: Massebalanse i den makroøkonomiske modellen MSG-EE. 1998. 49s. 110 kr inkl. mva. ISBN 82-537-4541-9

98/11 H. Bild, J.E. Finnvold, K.K. Lie, R. Nordhagen og A. Schjalm: Hvordan møter småbarnsfamiliene helsetjenesten? 1998. 99s. 115 kr inkl. mva. ISBN 82-537-4550-8

98/12 D. Roll-Hansen: Informasjonsteknologi i lærerutdanninga. 1998. 56s. 115 kr inkl. mva. ISBN 82-537-4554-0

98/13 A. Langørgen: Virkninger av lokalt bosettingsmønster på kostnader i kommunal tjenesteyting. 1998. 32s. 100 kr inkl. mva. ISBN 82-537-4555-9

98/14 Ø. Landfald og M. Bråthen: Evaluering av ordinære arbeidsmarkedstiltak: Dokumenta-sjon og analyse. 1998. 53s. 115 kr inkl. mva. ISBN 82-537-4561-3

98/15 T.I. Tysse og N. Keilman: Utvandring blant innvandrere 1975-1995. 1998. 160s. 155 kr inkl. mva. ISBN 82-537-4581-8

98/16 S. Blom: Levekår blant ikke-vestlige innvandrere i Norge. 1998. 81s. 115 kr inkl. mva. ISBN 82-537-4582-6

98/17 J. Epland: Endringer i fordelingen av husholdningsinntekt 1986-1996. 1998. 65s. 115 kr inkl. mva. ISBN 82-537-4584-2

98/18 K. Lund: Inntektsfordelinga i den norske landbruksbefolkninga og fordelingseffektar av direkte støtteordningar. 1998. 46s. 100 kr inkl. mva. ISBN 82-537-4585-0

98/19 H.K. Reppen: Bruk av folkebibliotek 1998. 1998. 46s. 115 kr inkl. mva. ISBN 82-537-4586-9

98/20 Ø. Landfald og M. Bråthen: Registerbasert evaluering av ordinære arbeidsmarkedstiltak 1996: Overgang til jobb og utdanning. 1998. 48s. 100 kr inkl. mva. ISBN 82-537-4596-6

98/21 J. Møen: Produktivitetsutviklingen i norsk industri 1980-1990 - en analyse av dynamikken basert på mikrodata. 1998. 85s. 115 kr inkl. mva. ISBN 82-537-4597-4

98/22 K. Flugsrud og G. Haakonsen: Utslipp til luft fra utenlandske skip i norske farvann 1996 og 1997. 1998. 37s. 100 kr inkl. mva. ISBN 82-537-4599-0

98/23 E. Nørgaard: The Norwegian Balance of Payments: Sources and methods. 1998. 72s. 115 kr inkl. mva. ISBN 82-537-4600-8

98/24 H. Hungnes: Imperfeksjoner i kapital-markedet. 1998. 37s. 100 kr inkl. mva. ISBN 82-537-4602-4

98/25 T. Løwe: Levekår i landbruket: En studie av landbruksbefolkningens levekår. 1998. 181s. 220 kr inkl. mva. ISBN 82-537-4603-2

99/1 A.C. Hansen: Fremskrivning av støybelastning for veitrafikk. 1999. 31s. 125 kr inkl. mva. ISBN 82-537-4659-8

99/2 T.W. Bersvendsen, J.L. Hass, K. Mork og B.H. Strand: Ressursinnsats, utslipp og rensing i den kommunale avløpssektoren, 1997. 1999. 71s. 140 kr inkl. mva. ISBN 82-537-4663-6

99/3 P. Boug: Modellering av faktoretterspørsel. 60s. 140 kr inkl. mva. ISBN 82-537-4665-2

B

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

Publikasjonen kan bestilles fra:

Statistisk sentralbyrå
Salg-og abonnementservice
N-2225 Kongsvinger

Telefon: 62 88 55 00
Telefaks: 62 88 55 95
E-post: salg-abonnement@ssb.no

eller:
Akademika – avdeling for
offentlige publikasjoner
Møllergt. 17
Postboks 8134 Dep.
N-0033 Oslo

Telefon: 22 11 67 70
Telefaks: 22 42 05 51

ISBN 82-537-4669-5
ISSN 0806-2056

Pris kr 125,00 inkl. mva.

Statistisk sentralbyrå
Statistics Norway

Falch Hurtigtrykk, Oslo