DR. SCIENT. EKSAMEN

Resamplingsmetodikk

VED: *Nils Lid Hjort*
FOR: *Magne Aldrin*
TID FOR EKSAMEN: *Onsdag 26. oktober 1988, fra kl. 13^{15} til ca. kl. 15^{30}*
FORBEREDELSE: *En uke, alle hjelpemidler*

Oppgavesettet inneholder to oppgaver, og er på tre sider. Kandidaten skal utarbeide løsning av så mange av deloppgavene som mulig, og gi en samlet fremstilling i forelesningsform på eksamen.

Oppgave 1

DET FØLGENDE DATAMATERIALE består av 25 uavhengige observasjoner $X_1, \ldots, X_{25}$ fra en felles, ukjent, underliggende fordeling F:

104.226	95.299	97.465	113.840	102.153	86.691	88.034
118.772	106.428	98.597	86.661	108.078	94.579	100.010
112.386	106.130	102.302	108.716	97.052	84.725	86.772
102.136	114.813	98.331	88.974			

Et histogram over data ser slik ut:

```
Midpoint   Count
  85       4  ****
  90       2  **
  95       4  ****
 100       6  ******
 105       3  ***
 110       3  ***
 115       2  **
 120       1  *
```

Alminnelige "summary statistics" (fra `Minitab`) er mean = 100.13, stdev = 9.72, median = 100.01, trmean = 99.99, semean = 1.94, min = 84.72, max = 118.77, Q1 = 91.78, Q3 = 107.25.

Vi skal anta at spredningsparameteren

$$\sigma = \sigma(F) = \sqrt{E_F(X - E_F X)^2} = \text{standardavviket til } F$$

er av særlig interesse, og skal se på punktestimater og konfidensintervall for denne. Et naturlig utgangspunkt er $\widehat{\sigma} = \sigma(\widehat{F})$, der $\widehat{F}$ er den vanlige empiriske fordelingsfunksjonen.

(a) Vis at $\widehat{\sigma}$ underestimerer σ i sin alminnelighet. Beregn $\widehat{\sigma}$ numerisk for det gitte datamateriale.

(b) Beregn jackknife-pseudoverdiene $\widehat{\sigma}_{(i)}$, samt estimatet $\widehat{\sigma}_{\text{JACK}}$ som prøver å rette opp for forventningsskjevhet.

(c) Skriv $E_F\widehat{\sigma} = \sigma(F) + b(F)$. Estimer skjevheten $b(F)$ ved bootstrapping, og beregn deretter det skjevhetsrettede punktestimatet

$$\widehat{\sigma}_{\text{BOOT}} = \widehat{\sigma} - \widehat{b}.$$

(d) Estimatoren $\widehat{\sigma}_{\text{BOOT}}$ som fremkom over har på sin side en eller annen systematisk skjevhet $c(F)$. Prøv å estimere denne, også ved bootstrapping, og beregn derved det dobbelt-rettede $\widehat{\sigma}_{\text{DOUBLEBOOT}}$.

Double, double, toil and trouble;
Fire burn and cauldron bubble.
— MacBeth, Shakespeare.

(e) Kan du tenke deg andre måter å "fjerne skjevheten" på?

(f) Man vil gjerne supplere punktestimatet for σ med et 90% konfidensintervall. Beregn noen slike: inkluder versjoner av "enkel bootstrap intervall", skjevhetsrettet (BC) intervall, og akselerasjon- og skjevhetsrettet (ABC) intervall. Kommenter forskjellene.

Oppgave 2

Biler synker i verdi etterhvert som de blir eldre. Datene (x_i, y_i) på neste side er dessverre simulerte, i mangel av ekte data, men pretenderer å være observasjonspar av x_i, bil nr. i's alder (regnet fra den dagen den ble solgt som ny), i år, og y_i, bilens verdi etter x_i år, regnet i prosent av nyverdi. La oss videre tenke oss at prisene er korrigert for inflasjon på passende måte, for å unngå visse tolkningsproblemer, og at biler som har vært ute for alvorlige kollisjoner ikke er med.

Vi skal konsentrere oss om biler med alder mellom 1 og 10 år. Den sanne, underliggende regresjonskurve er $y_0(x) = E(Y|x)$. Tidligere erfaring tilsier at en fornuftig parametrisk tilnærming er $y(x) = \alpha e^{-\beta x}$; se for eksempel Case Study #10.3 i Larsen & Marx' *An Introduction to Mathematical Statistics and Its Applications*, 1981. En rimelig statistisk beskrivelse av data er derfor

$$Y_i = \alpha e^{-\beta x_i}\,\varepsilon_i, \quad i = 1, \ldots, 30,$$

der de multiplikative korreksjonsfaktorene $\varepsilon_1, \ldots, \varepsilon_{30}$ er uavhengige og varierer rundt 1.

(a) Beregn estimater $\widehat{\alpha}$ og $\widehat{\beta}$ ved å bruke vanlig minste kvadraters metode på

$$\log Y_i = \log\alpha - \beta X_i + \text{støy}_i.$$

(b) Hvor meget har en fem år gammel bil sunket i verdi? La $\theta = E\{Y|X = 5\}$, og kommenter valget av $\widehat{\theta} = \widehat{y}(5) = \widehat{\alpha}\exp(-5\widehat{\beta})$ som estimator. Beregn den numeriske verdien.

(c) Hvor stor samplingfeil har estimatet $\widehat{\theta}$ for θ? La oss formalisere dette spørsmålet som et ønske om å estimere spredningsparameteren τ, der

$$\tau^2 = E(\widehat{\theta} - \theta)^2 = \text{Var}\,\widehat{\theta} + (E\widehat{\theta} - \theta)^2.$$

Estimer τ ved bootstrapping! Prøv gjerne forskjellige bootstrap-skjemaer. Et av dem skal være hel-parametrisk, ved å tilpasse ε_i'ene til en lognormal fordeling. Gjør rede for de forskjellige antagelser som gjøres for de forskjellige løsningsmetoder.

TABELL: *Observasjonspar* (x_i, y_i) *for 30 biler.*

x_i	y_i	x_i	y_i	x_i	y_i
1.56	47.6	4.54	29.3	6.03	19.8
1.90	62.1	4.65	22.4	6.15	14.4
1.92	47.3	4.75	23.0	6.23	15.7
1.99	53.7	5.05	26.0	6.43	15.9
2.02	40.7	5.31	18.7	6.65	17.9
2.26	47.9	5.38	15.6	6.89	9.7
2.93	38.5	5.56	19.4	7.86	9.4
3.16	48.4	5.61	16.5	8.16	10.7
4.26	25.0	5.91	14.0	9.10	5.4
4.34	25.6	5.95	17.9	9.36	6.0