UNIVERSITETET I OSLO

Gjennomsyn av Institutt

Ensaken i:	**STK 2000: Sentrale statistiske modeller og metoder**
Tid for Ensaken:	**Torsdag 15. desember 2005, kl. 14:30–17:30**
Hjelpemidler:	**«Formelsamling til STK 1100 og STK 1110», «Formelsamling til 1120», samt kalkulator**

Oppgavesettet har tre oppgaver og er på fire sider

Oppgave 1

Den italiensk-franske økonomiforskeren Vilfredo Pareto (1848–1923) arbeidet i Frankrike og Sveits. I samtiden var han kjent som en noe kranglete og ganske radikal politisk ulissprøver, men etterpåtiden husker ham best for hans arbeider i økonomisk og politisk økonomi, statistikk og statsvitenskap. Vi skal nå se på (visse modifikasjoner av) den såkalte Pareto-fordelingen, for hans Cours d'économie politique (1896–1897), og som ofte anvendes for å studere inntektsfordelingen i ulike stater.

(a) Vi ser at Y har en Pareto-fordeling med parameter θ (som skal > 0 positiv) hvis den kumulative fordelingsfunksjonen er på formen

$$F(y, \theta) = 1 - (1/y)^{\theta} \quad \text{for } y \geq 1.$$

Finn tettheten $f(y, \theta)$ samt modellens score-funksjon $U(y, \theta)$. Bruk generelle egenskaper til score-funksjonen og informasjonsmatrisen til å finne forventning og varians til variabelen $\log Y$.

(b) Anta nå uavhengige observasjoner $Y_1, \ldots, Y_{100}$ fra Pareto-fordelingen over vises seg å ha logaritmisk gjennomsnitt (1/100) $\sum_{i=1}^{100} \log Y_i = 3.33$. Finn maksimalsannsynlighetsestimatet $\hat{\theta}$, og lag så ville mente 95% konfidensintervall (i) for parameteren θ og (ii) for parameteren μ, medianen i Pareto-fordelingen.

(c) Vi skal så forestille oss at inntekten Y_i for personen i Oslo over 35 år, målt i antall tusen kroner pr. år, antas influert av kovariatene $x_{i,1}$ (antall års utdannelse etter grunnskolen) og $x_{i,2}$ (som er 0 for kvinner og 1 for menn), og at slike data $(Y_i, x_{i,1}, x_{i,2})$ foreligger for n individer $i = 1, \ldots, n$. For å analysere materialet bruker vi Pareto-modellen, der

$$Y_i \sim \text{Pareto}(\theta_i), \quad \text{med } \theta_i = \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}).$$

Kommenter hvor bredt du vil tolke parameteren β_2. Gjøra kjetter et enkelt gjett på størrelsen av β_0 og β_1, at du ikke alderdomlige konstellingen om årsgrunnlagen i Oslo. (Den Gulanges lagen Dyp presisjon, men kollen hvor bredt du kommer fem til ofra-rentten fo de to parametrene.)

(d) Sett opp en pesis formel for log-likelihood-funksjonen $\ell(\beta_0, \beta_1, \beta_2)$. Finn fremst en eksplisitt formel for den observerte informasjonsmatrisen $\hat{I}$. Kollen hvordan dette kan anvendes for å lage konfidensintervaller for modellens tre parametre.

(e) La $OK_{r\,r}$ ve modellen som bare tar hygde for utdannelsen u_1, og la $OK_{r\,r}$ ve den fulle modellen som tar med både utdannelse u_1 og kjønn u_2 som kovariater. Sett opp en eksplisitt formel for likelihoodforholdet $\Lambda(OK_r, OK)$, og kollen hvordan denne kan benyttes for å modere om det er avgjørende forskjell mellom menns og kvinners inntekt i Oslo, når det er korrigert for lengden av utdannelse. — ere om du kan vil eventuelle andre metoder for å avgjøre dette spørsmålet.

Oppgave 2

Vi skal besøke 150 nyfødte barn, om ikke i natura så i statistikk. Dette er gjorde ofrig det datasett som ble analysert og diskutert i forbindelse med Housets KOBig II, men med denne kvinnete anledning skal vi fem til et par andre spørsmålsstillinger. Datasettet består av 150 par, der i-te par med data for mor m. i er $(x_i, m_{i,0}, m_{i,2}, m_{i,3})$, der x_i = fødselsvekt (i 2500 gram), $m_{i,0}$ er kvinnens alder med barnets fødsel, $m_{i,2}$ er kvinnens vekt ikke fø graviditet, og $m_{i,3}$ er en indikator for om kvinnen røyker (1 om hun røyker og 0 om hun ikke røyker). Vi modellerer sannsynligheten for at et nyfødt barn skal vele minske enn 2500 gram ved

$$\pi_i = \pi(m_{i,0}, m_{i,2}, m_{i,3}) = \frac{e^{\ p(\beta_0 + \beta_0 m_{i,0} + \beta_2 m_{i,2} + \beta_3 m_{i,3})}}{1 + e^{\ p(\beta_0 + \beta_0 m_{i,0} + \beta_2 m_{i,2} + \beta_3 m_{i,3})}} = F(\beta_0 + \beta_0 m_{i,0} + \beta_2 m_{i,2} + \beta_3 m_{i,3})$$

for $i = 1, \ldots, 150$. Her er altså $F(x)$ den logistiske fordelingsfunksjonen $e^{\ p(x)}/(1 + e^{\ p(x)})$, og modellen svarer til almindelig logistisk regresjon, som behandlet i KOBig II.

(a) Den fødtsvemmelige og ikke-nyfødte vs. Dennes er 28 år gammel og veide 55 kg fø graviditet. Din oppgave er å finne et estimat $\hat{\pi}$ og et tilhørende 95% konfidensintervall for parameteren π, sannsynligheten for at akkurat denne barn skal vele minske enn 2500 gram, ut fra følgende lufen fra en R-kjøring:

```
m0 <- c(1,28,55,0)
plogis(t(m0) %*% coefficients %*% m0 )
0.0787
plogis(t(m0) %*% betahat )
-1.1954
```

Det opplyses om at `bettatt` er en 2 x 1-vektor med maksimum likelihood-estimatene for $\beta_0, \ldots, \beta_3$, mens `Cntatt` er forventet til den observerte informasjonsmatrise.

(b) Så skal vi undersøke om det er vesentlig forskjell mellom de slettede og de stryppede modeller, med hensyn til den logistiske regresjonsmodellen og dennes parametre. Vår felles datamaterialet i en `babysmall`-fil, for de $n_0 = 93$ mødre som veide mindre enn 55.0 kg før graviditet, og en `babybig`-fil, for de $n_1 = 93$ mødre som veide mer enn 55.0 kg før graviditet. Deretter har jeg anvendt logistisk regresjon for hver av de to kalkulasjonene, og benyttet maksimum likelihood-estimatene

```
bettatt1 <- glm(ysmall ~ xsmall, family=binomial)$coef
bettatt2 <- glm(ybig ~ xbig, family=binomial)$coef
```

samt observerte informasjonsmatriser `Cntatt1` og `Cntatt2` for disse. Dette gir videre følgende R-kode:

```
bettadiff <- bettatt2 - bettatt1
mat1 <- solve(Cntatt1)
mat2 <- solve(Cntatt2)
Wn <- t(bettadiff) %*% solve(mat1 + mat2) %*% bettadiff
pchisq(Wn)
0.169
```

Bortsett fra disse opplysningene til å støtte den hypotese at det ikke er forskjell mellom de to nærliggende parametervektorene for de to logistiske regresjonene. Forklar disse argumenter og gjør rede for de avgjørende resultater du trenger å benytte til. Om du har tid kan du også antyde eventuelle andre metoder som kunne komme til anvendelse for å teste likheten mellom de to regresjonene.

Oppgave 3

Den er hasardrisiko å denne, men vi har vel ikke noe valg: Vi må sette opp det øyeblikksbilde, det vil si det øyeblikk de er ødelagt, som Shakespeare sier (i Ved det ene).

(a) I et bestemt laboratorium-eksperiment utsettes disse basiller for en type påkjenning, i et kontrollert kammer, som gjør at de dør etter en stokastisk tid T, den basiliotiske forklaringen til T er $\{T\} = 1 + \log(y) + \varepsilon$. Basillene dør uavhengig av hverandre, og tilsvarende en i timen. Et stort antall basiller slippes nå ut i kammeret. Hvor stor bøtteel av basillene kan forventes å overleve én time, og hvor stor bøtteel kan forventes å fortsatt å leve i lite etter to timer? Etter hvor lang tid kan man regne med at halvparten av basillene er døde?

(b) Gen et ligningssett med ikke-lineært eksponentielt utslag det at Bernoulliforsøkene på formen $(x, \theta) = 1$ hvis for $x \geq 0$, hvor θ er en ukjent positiv parameter. Vi har sanns lun av typen $(x_0, y_0), \ldots, (x_n, y_n)$ for n uavhengige forsøk, hvor $y_i = 1$ om forsøksutfallene er blitt registrert og $y_i = 0$ om forsøkene er sensurert. Sett opp et uttrykk for likelihood-funksjonen $L(\theta)$ for data.

(c) I stimasjonen for punkt (b), finn en eksplisitt formel for et maksimalsannsynlighets-funksjon likelihood-estimatoren $\hat{\theta}$. Finnes det flere data der det blir delt på å finne et tilsvarende 95% konfidensintervall for parameteren θ.

(d) For stimasjonen fra (b)–(c), foreslå en passende metode for å sjekke om den underliggende modellantagelsen om Bernoulliforsøk er korrekt.

(e) Vi finner nå modellen i (b)–(d), og arbeider med det følgende scenario: Basilen lever og dø uavhengig av hverandre, og hver enkelt basill har sin konstante Bernoulliprøve, men med forskjellige parametre for forskjellige basiller. Vi gir basill i altså Bernoulliprøve på formen $(x, \theta) = \theta$ for $x \geq 0$ [og hvor i følgende overlevelses-funksjon $S(x, \theta) = e^{-\theta x}$], men θ-ene varierer i populasjonen, etter en fordeling som vi har skal vite en standard eksponensiell, altså med tettheten $f(\theta) = e^{-\theta}$ for $\theta \geq 0$.

Hva er nå sannsynligheten for at en tilfeldig valgt basill skal overleve åtte timer? Hva blir selve observerte Bernoulliprøven, altså Bernoulliprøven for en tilfeldig valgt basill? Svaret skal bli en antagelse funksjon av tiden x. Kommentere hvor gunstig til at den observerte Bernoulliprøve blir antagelse selv om hvert enkelt individs Bernoulliprøve er konstant.