ANDRE OBLIGATORISKE OPPGAVE
STK2120 VÅREN 2011

Innleveringfrist: Torsdag 14. april kl 14.30 i obligkassa i 7. etasje i Niels Henrik Abels hus. Erfaringsmessig blir det lange køer rett før innleveringsfristen, så det er smart å levere tidligere. Husk å bruke den offisielle obligforsiden ved innlevering! Dersom du på grunn av sykdom eller lignende har behov for å utsette innleveringen, må du sende søknad til Robin Bjørnetun Jacobsen (rom B718, NHA, e-post: studieinfo@math.uio.no, tlf. 22 85 59 07). Husk at sykdom må dokumenteres ved legeattest! Se forøvrig
http://www.math.uio.no/academics/obligregler.shtml
for nærmere informasjon om reglement rundt obligatoriske oppgaver ved matematisk institutt.

Oppgaven er obligatorisk og studenter som ikke får besvarelsen godkjent, vil ikke få adgang til avsluttende eksamen. For å få besvarelsen godkjent, må man ha gjort et forsøk på å løse alle deloppgaver. Du kan få poeng på en oppgave selv om du ikke har kommet frem til et svar, og det er derfor viktig at du leverer inn det du har kommet frem til. Det er lov å samarbeide og bruke alle slags hjelpemidler. Den innleverte besvarelsen skal imidlertid være skrevet av deg (for hånd eller på datamaskin) og gjenspeile din forståelse av stoffet. Er vi i tvil om du virkelig har forstått det du har levert inn, kan vi be deg om en muntlig redegjørelse.

Foruten svar på de spesifikke spørsmål skal du også legge ved alle relevante plott og utskrifter av kommandoer du bruker. I den skriftlige besvarelsen skal du *forklare* hvordan de enkelte punktene er løst, og du skal *diskutere* de resultatene du kommer fram til. Det er valgfritt om du vil skrive besvarelsen for hånd eller om du vil bruke et tekstbehandlingsprogram. Uansett skal resultatene av kjøringer tas med i besvarelsen på en hensiktsmessig måte.

Oppgave 1
I notatet om numerisk optimering så vi på logistisk regresjon for å analysere dødelighet av biller relatert til dose av gift. Dataene var gitt i tabellform (Tabell 8.2), mens algoritmen som ble implementert var relatert til data på individnivå. Vi skal i denne oppgaven se på hvordan vi kan analysere slike tabelldata mer direkte.

Anta nå n_i er antall biller som er gitt giftnivå x_i og at Y_i er antall biller som dør. La N være antall ulike giftnivåer. Da vil vi nå anta $Y_1, ..., Y_N$ er uavhengige og

$$Y_i \sim \text{binom}(n_i, p(x_i, \boldsymbol{\beta}))$$

der

$$p(x_i, \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

For å gjøre det notasjonsmessig litt enklere i det etterfølgende vil vi skrive

$$p_i = p(x_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}$$

der $\boldsymbol{x}_i = (x_{i0}, x_{i1})^T = (1, x_i)^T, \boldsymbol{\beta} = (\beta_0, \beta_1)^T$.

(a) Plot y_i/n_i mot x_i. Hvorfor er dette et rimelig plot for å se på hvordan y_i varierer med x_i?

(b) Vis at log-likelihood funksjonen i dette tilfellet kan skrives som

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} [y_i \log(p_i) + (n_i - y_i) \log(1 - p_i)].$$

(c) Vis at det k-te elementet i skår funksjonen generelt kan skrives som

$$s_k(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_k} l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{y_i - n_i p_i}{p_i(1 - p_i)} \frac{\partial}{\partial \beta_k} p_i$$

(d) Vis at

$$\frac{\partial}{\partial \beta_k} p_i = p_i(1 - p_i) x_{ik}$$

og sett det inn i uttrykkene for skårfunksjonen for å gi et forenklet uttrykk for skårfunksjonen.

(e) Finn også den observerte informasjonsmatrisen og vis at den samsvarer med likning (8.17) i notatet om numerisk optimering hvis $n_i = 1$ for alle i.

Vis at den observerte informasjonsmatrisen alltid er positiv definitt.

(f) Implementer en Newton-Raphson algoritme som maksimerer $l(\boldsymbol{\beta})$ ved å modifisere algoritmen gitt i Figur 8.3 i notatet om numerisk optimering (koden er tilgjengelig fra kursets hjemmeside). Merk at du i tillegg til x_i, y_i nå også må angi n_i.

(g) Kjør algoritmen og sikre deg om at du får de samme resultater som gitt i notatet.

(h) I plottet fra deloppgave (a), legg på en linje som viser $p(x_i, \widehat{\boldsymbol{\beta}})$ som en funksjon av x_i. Kommentér figuren.

Beregn også $l(\widehat{\boldsymbol{\beta}})$ og lagre denne. Vi skal bruke denne verdien i neste oppgave for å sammenlikne ulike modeller.

I notatet om numerisk optimering ble asymptotiske ("large sample") egenskaper for $\widehat{\boldsymbol{\beta}}$ brukt for å si noe om egenskaper til estimatene. Vi vil her se på bruk av *bootstrap* metoder i stedet. I vår situasjon har vi et sett av bivariate observasjoner $\{(x_i, y_i), i = 1, ..., N\}$ og det er ikke helt opplagt hvordan man skal generere bootstrap simulerte data $\{(x_i^*, y_i^*), i = 1, ..., N\}$ i dette tilfellet. Vi vil se på en enkel *parametrisk bootstrap* prosedyre der vi lar

$$\begin{aligned} x_i^* =& x_i \\ y_i^* \sim& \text{binom}(n_i, p(x_i, \widehat{\boldsymbol{\beta}})) \end{aligned}$$

(i) Diskutér denne typen simulering i forhold til den metode som er diskutert i bootstrap notatet, avsnitt 6.

(j) Simuler $B = 1000$ bootstrap utvalg, finn ML-estimatet $\widehat{\boldsymbol{\beta}}^*$ for hvert utvalg.

Hint: Det kan være lurt å her bruke $\widehat{\boldsymbol{\beta}}$ som start-verdier for algoritmen.

Bruk dette til å estimere forventningsskjevhet og standard feil for estimatene samt for å lage 95% konfidensintervall for β-ene.

Sammenlikn med resultatene en får basert på en asymptotisk tilnærming.

(k) Som nevnt i bootstrap notat, kan en gjøre simultan analyse av flere estimater/parametre samtidig ved bootstrapping. La nå $p_i = p(x_i, \boldsymbol{\beta})$ for en gitt x_i verdi være en parameter av interesse. Forklar hvordan du kan lage et 95% konfidensintervaller for p_i basert på dine B bootstrap simuleringer.

(l) Lag 95% konfidensintervall for $p_i, i = 1, ..., N$ der x_i tar verdiene gitt i datasettet. I plottet ditt fra deloppgave (h), legg til en linje som viser nedre grenser for konfidensintervallene for p_i og en linje for øvre grense.

Oppgave 2

Vi skal i denne oppgaven se på en alternativ modell for bille dataene. Anta nå

$$Y_i \sim \text{binom}(n_i, p_i)$$

der

$$p_i = p(x_i, \boldsymbol{\beta}) = 1 - \exp(-\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}))$$

(a) Vis at det k-te elementet i skårfunksjonen i dette tilfellet er gitt ved

$$s_k(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{y_i - n_i p_i}{p_i} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) x_{ik}$$

En kan også vise at Fisher informasjonsmatrisen har element (kl) lik

$$\bar{I}_{kl} = \sum_{i=1}^{N} n_i \frac{1-p_i}{p_i} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) x_{ik} x_{il}$$

men dette behøver du ikke å vise.

(b) Implementer en Fisher-skår algoritmen for denne modellen.

Kjør algoritmen for å finne estimatene.

(c) I figuren du laget i oppgave 1, legg nå til en linje som viser $p(x_i, \widehat{\boldsymbol{\beta}})$ som en funksjon av x_i for denne modellen. Kommentér resultatene.

Beregn også $l(\widehat{\boldsymbol{\beta}})$ og sammenlikn med verdien du fikk for modellen i oppgave 1. Basert på plottet og log-likelihood-verdiene, hvilken modell vil du foretrekke?

(d) Gjenta deloppgavene $(j) - (l)$ fra oppgave 1, nå for den alternative modellen.

Kommentér resultater.