2. obligatoriske oppgave i STK3100/STK4100
Høst 2007

Utlevering: Fredag 2. november

Innleveringsfrist: Fredag 16. november, kl. 14:30

Besvarelsen innleveres ved ekspedisjonen i 7. etasje, Niels Henrik Abels hus

Dette er det andre settet med obligatoriske innlevering for STK31000 og STK4100 høsten 2007. Oppgavesettet består av to oppgaver. Det er valgfritt om du vil skrive besvarelsen for hånd eller om du vil bruke et tekstbehandlingsprogram. Der du bruker `R` (eller et annet program), må utskrifter legges ved eller limes inn. Hvis flere studenter samarbeider om å løse oppgavene, må likevel hver student levere sin selvstendige besvarelse. Det må gå fram av besvarelsen hvem du har samarbeidet med. Se ellers "Regelverk for obligatoriske oppgaver" som er gitt på kursets hjemmeside.

Oppgave 1

Hvorfor er det et mange forskjellige fuglearter på noen øyer, men få andre steder? Ornitologer har telt opp antall fuglearter y på hver av 14 øyer utenfor Ecuador. Mer spesifikt handler det om antall arter funnet på visse paramoer på disse øyene (en paramo er en bestemt type høytliggende platå, typisk i tropiske områder av Syd-Amerika). Tabellen under gir y sammen med AR (areal, regnet pr. tusen kvadratkilometer), EL (et representativt elevation-tall, altså høyde over havet for vedkommende paramo, i kilometer), DEC (distanse til Ecuador, i kilometer), DNI (distanse til nærmeste øy, i kilometer):

y	AR	EL	DEC	DNI
36	0.33	1.26	36	14
30	0.50	1.17	234	13
37	2.03	1.06	543	83
35	0.99	1.90	551	23
11	0.03	0.46	773	45
21	2.17	2.00	801	14
11	0.22	0.70	950	14
13	0.14	0.74	958	5
17	0.05	0.61	995	29
13	0.07	0.66	1065	55
29	1.80	1.50	1167	35
4	0.17	0.75	1182	75
18	0.61	2.28	1238	75
15	0.07	0.55	1380	35

Disse dataene finnes også på hjemmesiden for STK3100.

a) La Y_i være antall fuglearter på vedkommende paramo på øy nr. i. Vi ser på Y_i-ene som resultat av uavhengige Poisson-variable med parametre $\mu_1, \ldots, \mu_{14}$, der $\mu_i = \exp(\beta_1 + \beta_2 AR_i + \beta_3 EL_i + \beta_4 DEC_i + \beta_5 DNI_i)$ for $i = 1, \ldots, 14$. Diskuter kort hva modellen innebærer, og om den synes rimelig.

b) Sett opp en eksplisitt formel for log-likelihood-funksjonen for denne modellen. Finn også en formel for informasjonsmatrisen, $\mathcal{J}(\beta)$. Hvorfor er observert og forventet informasjonsmatrise like for denne modellen?

c) Bruk glm-rutinen i R til å tilpasse modellen i (a): finn parameterestimater og standardavvik.

d) Hvilke av de fire faktorene AR, EL, DEC, DNI synes spesielt viktige for utbredelsen av fuglearter på en tropisk paramo, og hvilke faktorer synes ikke så viktige?

e) Test hypotesen at paramoens areal ikke har direkte betydning for antall fuglearter. Gi en konklusjon.

f) Test så hypotesen at hverken EL eller DNI har direkte innflydelse for fugleartenes mangfold.

g) Ornitologenes budsjett strakk ikke til for å undersøke en bestemt øy nr. 15, som har areal 1100 kvadratkilometer, elevation-tall 800 meter, er 666 kilometer fra Ecuador, og er 22 kilometer fra nærmeste store øy. Med Y = antall fuglearter på denne øyas paramo gi et estimat for $\mathrm{E}[Y]$. Gi også en rimelig nedre og øvre konfidensgrense for $\mathrm{E}[Y]$.

h) Undersøk om Poisson-modellen i (a) synes å passe godt for de aktuelle data.

i) Prøv å komme frem til en god "endelig modell" for Y_i-ene.

Oppgave 2

I denne oppgaven skal vi nærmere på dataene fra kasus-kontroll studien i Oppgave 22, sammenfattet i følgende tabell:

Kjønn	Gruppe	Ektefelle røyker	Ektefelle røyker ikke	Total
Mann	Case	2	6	8
	Control	26	154	180
	Total	28	160	188
Kvinne	Case	14	8	22
	Control	61	72	133
	Total	75	80	155

Vi skal bare se på dataene for **menn**. Et problem med disse er at det kun var 2 syke som var utsatt for passiv røyking fra ektefelle. Dermed er det usikkert om parameterestimatenes tilnærming normalfordeling holder og dermed usikkerhet mht. utførte tester og beregnede konfidensintervaller. Vi skal alternativt finne konfidensintervaller basert på likelihood ratio tester (LRT).

a) Analyser pånytt dataene (for menn) med logistisk regresjon. Beregn 95% konfidensintervall for odds-ratioen (OR) mellom dem som er og ikke er utsatt for passiv røyking. Test også nullhypotesen

$$\mathrm{H}_0 : \mathrm{OR} = 1$$

både med Wald-test og med LRT.

b) Test også nullhypotesen

$$\mathrm{H}_0 : \mathrm{OR} = 3$$

både med Wald-test og med LRT.

c) For LRT-testen i punkt b) måtte du benytte `offset` som beskrevet i et forelesningsnotat. Forklar hva som ligger i dette begrepet og hvordan dette gir LRT.

d) Forklar hvorfor

$$\{\beta_{10} : \ \mathrm{H}_0 : \mathrm{OR} = \exp(\beta_{10}) \text{ ikke forkastes ved 5\% nivå } \}$$

er et tilnærmet 95% konfidensintervall for $\beta_1 = \log(\mathrm{OR})$.

e) Gjennomfør testen for en passende mengde av verdier av β_{10}. Plott residualdeviansene mot valgte verdier av β_{10} eller $\mathrm{OR} = \exp(\beta_{10})$ og finn fra plottet 95% LR-intervall for OR. Sammenlign med Wald-intervallet i punkt a).

f) Anta nå at antall eksponerte blant de syke er $A = 1$ istedetfor $A = 2$ som i tabellen, men at dataene forøvrig er de samme. Gjør de tilsvarende analysene som i punktene a)-e) og sammenlign resultatene.

g) Anta så at antall eksponerte blant de syke er $A = 0$ mens at dataene forøvrig er de samme. Gjør pånytt tilsvarende analyser som i punktene a)-e). Kommenter.