UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK3100 — Innføring i generaliserte lineære modeller
Eksamensdag:	Mandag 6. desember 2010
Tid for eksamen:	14.30 – 18.30
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabell over normalfordeling og χ^2-fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK1120/STK2120

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1

I vedlegg 1 er det et datasett fra 35 operasjoner der forekomsten av sår hals etter narkose er registrert. Her betrakter vi sår hals (`sore`) som respons, 0 svarer til nei og 1 til ja. Kovariatene er lengden på operasjonen i minutter (`duration`) og to typer (`type`) utstyr brukt til å holde luftveiene åpne under operasjonen.

Utskriften nedenfor er basert på en modell der responsen er binomisk fordelt, $Bin(m, \pi)$, der m er lik 1, altså det som kalles binære responser. Linkfunksjonen er logit link. I første omgang skal vi bare betrakte varighet som kovariat, dvs. prediktoren har formen

$$\eta = \beta_0 + \beta_1 x$$

der x er lengden på operasjonen.

```
Call:
glm(formula = sore ~ I(duration), family = binomial, data = sore)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0964  -0.7392   0.3020   0.8711   1.3753

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.21358    0.99874  -2.216  0.02667 *
I(duration)  0.07038    0.02667   2.639  0.00831 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.180  on 34  degrees of freedom
Residual deviance: 33.651  on 33  degrees of freedom
AIC: 37.651
```

(Fortsettes på side 2.)

a) Forklar hvordan en generalisert lineær modell er definert, og hvorfor modellen ovenfor er av denne typen.

b) Estimer oddsforholdet for forekomst av sår hals mellom to operasjoner der den ene varer 30 og den andre 40 minutter. Angi også et 95% konfidensintervall for dette oddsforholdet.

c) Hva er den predikerte sannsynligheten for sår hals ved en operasjon som varer 40 minutter? Beregn også et 95% konfidensintervall. Her trenger du å vite at den estimerte korrelasjonen mellom $\hat{\beta}_0$ og $\hat{\beta}_1$ er -0.906.

d) I deviansanalyse-tabellen nedenfor finner du deviansen for modeller som også inneholder kovariaten `type` og et kvadratisk ledd i `duration`. Antallet frihetsgrader fjernet. Fyll ut de manglende tallene. Begrunn deretter at modellen vi så på punktene a)-c) er et rimelig valg. Du kan anta at den mest generelle modellen (`Model 4`) har en tilfredsstillende tilpasning.

```
Analysis of Deviance Table

Model 1: sore ~ 1
Model 2: sore ~ I(duration)
Model 3: sore ~ I(duration) + factor(type)
Model 4: sore ~ I(duration) + I(duration^2) + factor(type)
  Resid. Df Resid. Dev Df Deviance
1        ?     46.180
2        ?     33.651  ?   12.528
3        ?     30.138  ?    3.513
4        ?     30.133  ?    0.005
```

e) La y_i og $\hat{\pi}_i, 1 = 1, \ldots, 35$ være henholdsvis de observerte og de tilpassede responsverdiene. Vis at deviansen for binomiske modeller med binær respons kan skrives

$$-2\sum_{i=1}^{35}[\hat{\pi}_i \log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) + \log(1-\hat{\pi}_i)].$$

Forklar omhyggelig hvorfor det medfører at deviansen er uegnet som føyningsmål i dette tilfellet.

Oppgave 2

Tabellen nedenfor er et berømt datasett fra en undersøkelse om sammenhengen mellom røyking og dødsfall på grunn av hjertesykdommer blant britiske leger. Antall dødsfall er respons og alder (`age`) og røyking (`smoker`) er kovariater. Vi lar alder være en numerisk variabel, og tilordner verdiene, eller skårene, 40, 50, 60, 70 og 80 til de fem aldersgruppene. Røyking er en faktor med to nivåer der 0 betegner ikke-røyker og 1 røyker. I tillegg er det registrert en variabel (`persyear`) som angir antall leveår i de ulike kategoriene.

(Fortsettes på side 3.)

Tabell 1: Dødelighet og røyking.

	Personår		Hjerterelatert dødsfall	
Alder	Ikke-røyker	Røyker	Ikke-røyker	Røyker
35-44	18793	52407	2	32
45-54	10673	43248	12	104
55-64	5710	28612	28	206
65-74	2585	12633	28	186
75-84	1462	5317	31	102

Utskriften viser resultatet av en tilpasning av en modell der responsen har en Poisson-fordeling og det er benyttet en kanonisk log-link.

```
Call:
glm(formula = deaths ~ offset(log(persyear)) + I(age) + I(age^2) +
    factor(smoker) + I(age):factor(smoker), family = poisson,
    data = coro)


Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.971e+01  1.253e+00 -15.734  < 2e-16 ***
I(age)                     3.565e-01  3.631e-02   9.819  < 2e-16 ***
I(age^2)                  -1.978e-03  2.736e-04  -7.228 4.89e-13 ***
factor(smoker)1            2.370e+00  6.559e-01   3.613 0.000303 ***
I(age):factor(smoker)1    -3.084e-02  9.699e-03  -3.180 0.001474 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 936.6589  on 9  degrees of freedom
Residual deviance:   1.6661  on 5  degrees of freedom
AIC: 66.734
```

a) Forklar hvorfor antagelsen om Poisson-fordelte responser er rimelig i denne situasjonen. Gi en eksplisitt beskrivelse av hvordan kovariatene inngår i modellen som er tilpasset i utskriften ovenfor. Kommenter resultatet.

b) Forklar hva offset er. Hvorfor er det rimelig å benytte offset i dette tilfellet?

c) Uttrykk betydningen av røyking for denne typen dødelighet ved relevante forhold mellom ratene (rate ratios). Beregn spesielt forholdene mellom ratene for røykere og ikke-røykere for leger som er 40 år og for leger som er 70 år. Diskuter resultatet.

(Fortsettes på side 4.)

Nedenfor finner du utskriften for tilpasning av en mer generell modell.

```
Call:
glm(formula = deaths ~ offset(log(persyear)) + I(age) + I(age^2) +
    factor(smoker) + I(age):factor(smoker) + I(age^2):factor(smoker),
    family = poisson, data = coro)


Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.153e+01  3.197e+00  -6.736 1.63e-11 ***
I(age)                       4.148e-01  1.004e-01   4.130 3.62e-05 ***
I(age^2)                    -2.430e-03  7.739e-04  -3.140  0.00169 **
factor(smoker)1              4.445e+00  3.391e+00   1.311  0.18991
I(age):factor(smoker)1      -9.755e-02  1.069e-01  -0.912  0.36160
I(age^2):factor(smoker)1     5.196e-04  8.273e-04   0.628  0.52999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 936.6589  on 9  degrees of freedom
Residual deviance:   1.2623  on 4  degrees of freedom
AIC: 68.33

Number of Fisher Scoring iterations: 4
```

d) Vi ser at to siste estimatene som beskriver samspill mellom alder og røyking ikke er signifikante hver for seg. Utfør en Wald test for å teste hypotesen om at de tilsvarende koeffisientene er lik null samtidig, dvs den simultane hypotesen at begge er null. Matrisen nedenfor er den estimerte kovariansmatrisen for estimatorene til de to koeffesientene.

```
                            I(age):factor(smoker)1  I(age^2):factor(smoker)1
I(age):factor(smoker)1        1.143363e-02              -8.807653e-05
I(age^2):factor(smoker)1     -8.807653e-05               6.844424e-07
```

Den inverse kovariansmatrisen har diagonalelementer `10038.02` og `16768.5354e+04`. Elementene utenfor diagonalen er `12917.2852e+02`.

e) Forklar hvorfor den forventede og observerte informasjonsmatrisen blir like i modeller av den typen vi har sett på i denne oppgaven.

SLUTT

(Fortsettes på side 5.)

Vedlegg 1

	duration	type	sore
1	45	0	0
2	15	0	0
3	40	0	1
4	83	1	1
5	90	1	1
6	25	1	1
7	35	0	1
8	65	0	1
9	95	0	1
10	35	0	1
11	75	0	1
12	45	1	1
13	50	1	0
14	75	1	1
15	30	0	0
16	25	0	1
17	20	1	0
18	60	1	1
19	70	1	1
20	30	0	1
21	60	0	1
22	61	0	0
23	65	0	1
24	15	1	0
25	20	1	0
26	45	0	1
27	15	1	0
28	25	0	1
29	15	1	0
30	30	0	1
31	40	0	1
32	15	1	0
33	135	1	1
34	20	1	0
35	40	1	0