UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST 202 — Statistiske slutninger for den eksponentielle fordelingsklasse.
Eksamensdag:	Fredag 15. desember 1995.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 7 sider.	
Vedlegg:	Tabeller for χ^2, t og normal fordeling
Tillatte hjelpemidler:	Formelsamlinger for ST 101, ST 102, ST 103, lommeregner.

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

Oppgave 1.

Vi vil i denne oppgaven studere skader på skip forårsaket av bølger. En er interessert i risikoen for skader i forhold til 4 kovariater:

aggr.serv	=	Antall måneder i drift
type	=	Skipstype A,B,C,D eller E
y.constr	=	Bygningsår 1960-64, 1965-69, 1970-74 eller 1975-79
p.oper	=	Operasjonsperiode: 1960-74, 1975-79

a) Vi vil starte med å anta modellen

log (forventet antall skader) =

β_0 + log(antall måneder i drift)
+ (effekt av skipstype)
+ (effekt av bygningsår)
+ (effekt av operasjonsperiode)

Hvilke antagelser ligger i denne modellen?

Er det rimelig å estimere regresjonskoeffisienter for alle variable?

(Fortsettes side 2.)

b) Hva slags typer variable kan de ulike kovariatene være?
Sett opp antall parametre som inngår i modellen for hver variabel.
I tilfeller der kovariatene kan anta flere ulike typer, diskuter fordeler og ulemper med de ulike valg.

Det er rimelig å anta at responsen er Poisson fordelt. Vi vil forutsette dette i det følgende.

c) Skriv opp definisjonen på generaliserte lineære modeller og vis at vår problemstilling faller innenfor denne rammen. Hva blir link funksjonen?

d) Gjør rede for devians-begrepet.
Hva kan vi bruke deviansen til?
Hva blir deviansen for Poisson-modellen?

e) I tabellen nedenfor er utskrift av tilpasning basert på modellen i a). med *type, y.constr* og *p.oper* som faktorer og koeffisienten foran log (antall måneder i drift) satt til 1.

Utfra disse tallene, gi en vurdering av kovariatenes betydning.
Gi en tolkning av regresjonskoeffisienten for *p.oper*.
(Kontrasten "treatment" er brukt under tilpasning.)

```
Coefficients:
                  Value   Std. Error     t value
(Intercept)  -6.4059016    0.2174315  -29.461700
      typeB  -0.5433443    0.1775869   -3.059596
      typeC  -0.6873773    0.3281646   -2.094611
      typeD  -0.0759614    0.2905555   -0.261435
      typeE   0.3255795    0.2358758    1.380300
  y.contr65   0.6971404    0.1496286    4.659139
 y-constr70   0.8184266    0.1697504    4.821354
 y.constr75   0.4534266    0.2331490    1.944793
     p.oper   0.3844670    0.1182584    3.251077

    Null Deviance:  146.3283 on 33 degrees of freedom
Residual Deviance:  38.69505 on 25 degrees of freedom
```

f) En utvidet modell ville være å inkludere interaksjonsledd mellom kovariatene. Vi vil undersøke om dette er en fornuftig modell. Nedenfor er en deviansanalysetabell for dataene angitt.

```
                  Df  Deviance Resid.  Df  Resid.  Dev
            NULL                       33     146.3283
            type   4  55.43906         29      90.8893
       y.constr    3  41.53409         26      49.3552
         p.oper    1  10.66014         25      38.6951
  type:y.constr   12  24.10231         13      14.5927
    type:p.oper    4  6.06979          9       8.5230
y.constr:p.oper    2  1,66621          7       6.8567
```

Forklar de ulike kolonner i tabellen. Hvilken modell ville du velge? Begrunn svaret.

(Fortsettes side 3.)

g) Forklar hva som menes med over-dispersjon. Diskuter mulige årsaker for over-dispersjon i det konkrete problemet.

For modellen der kun interaksjonsledd mellom *type* og *y.constr* er inkludert i tillegg til hovedeffektene, ble dispersjonsparameteren estimert til $\hat{\phi} = 1.336$. Er over-dispersjonen signifikant?

Oppgave 2.

I et studie ved Stanford Clinical Research Center ønsket en å se på sammenhengen mellom kjemiske subkliniske prøver og diabetes. 145 voksne personer ble undersøkt. Responsen, Y, kunne anta 3 verdier:

$$Y = \begin{cases} 1 & \text{for personer uten diabetes} \\ 2 & \text{for kjemisk diabetes} \\ 3 & \text{for åpen diabetes} \end{cases}$$

For hver person ble videre følgende variable registrert:

$$x_1 = \text{Insulinareal}$$
$$x_2 = \text{Insulinresistanse}$$

En ønsker å undersøke om disse variablene kan benyttes for å karakterisere personer i forhold til diabetes.

a) En vanlig modell for slike data er

$$\pi_j(x) = \frac{\exp(\eta_j(\boldsymbol{x}))}{\sum_{l=1}^{k} \exp(\eta_l(\boldsymbol{x}))}$$

der $\pi_j(\boldsymbol{x}) = Pr(Y = j|\boldsymbol{x})$ og $\eta_j(\boldsymbol{x}) = \beta_{j,0} + \sum_{k=1}^{p} \beta_{j,k} x_k$ hvor p er antall kovariater.

Er denne modellen innenfor den eksponensielle klasse? (Begrunn svaret). Hvorfor vil en ofte sette $\eta_1(\boldsymbol{x}) = 0$?

b) Sett opp likelihood-funksjonen og regn deg frem til likelihood-likningene til bestemmelse av $\hat{\boldsymbol{\beta}}_s = (\widehat{\beta}_{j,0}, \cdots, \widehat{\beta}_{s,p})$, $j = 1, 2, 3$, sannsynlighetsmaksimeringsestimatene for $\beta_{j,k}$-ene.

Beskriv kort en numerisk algoritme for å beregne estimatene.

(Fortsettes side 4.)

c) Beregn sannsynlighetene

$$Pr(Y = j | Y \in \{1, j\}, \boldsymbol{x}) \quad j = 2, 3$$

Hva slags modell tilsvarer dette?
Hvordan kan vi utnytte dette ved estimering?
Diskuter fordeler og ulemper ved denne prosedyren.

d) Tabellene nedenfor viser resultatene av tilpasninger av dataene

$$\{(x_{i,1}, x_{i,2}, y_i), y_i \in \{1, 2\}\}$$

og

$$\{(x_{i,1}, x_{i,2}, y_i), y_i \in \{1, 3\}\}$$

henholdsvis

Tabell 1

```
Coefficients:
                    Value   Std. Error    t value

                    Value   Std. Error    t value
(Intercept)    3.32189801  1.707236389   1.945775
         x1    0.01627855  0.004757766   3.421468
         x2   -0.02590072  0.008058611  -3.214043

    Null Deviance:  95.52384 on 68 degrees of freedom
Residual Deviance:  39.74526 on 66 degrees of freedom
```

Tabell 2

```
Coefficients:
                    Value   Std. Error    t value

                    Value   Std. Error    t value
(Intercept)    6.06194231  1.529096338   3.964395
         x1    0.01219343  0.004673021   2.609325
         x2   -0.03655439  0.008211284  -4.451726

    Null Deviance:  133.6729 on 108 degrees of freedom
Residual Deviance:  34.84639 on 106 degrees of freedom
```

Forklar hva de ulike kolonner betyr. Anta

$$x_1 = 124$$
$$x_2 = 55$$

Beregn $\pi_j(\boldsymbol{x})$ for $j = 1, 2, 3$.

En alternativ fremgangsmåte for analyse av slike data er å modellere hvordan kovariatene X_1 og X_2 varierer gitt Y.
For enkelthets skyld vil vi nå kun betrakte én forklaringsvariabel, X_1.

(Fortsettes side 5.)

e) Anta $\Pr(Y = j) = \pi_j$, $j = 1, 2, 3$ og $x_1|Y = j \sim N(\mu_j, \sigma_j^2)$. Vis at

$$\Pr(Y_j|x_1) = \frac{\exp(\eta_j(x_1))}{1 + \sum_{l=2}^{3} \exp(\eta_l(x_1))} \qquad j = 2, 3$$

der $\eta_j(x_1) = \alpha_j + \beta_j x_1 + \gamma_j x_1^2$ og

$$\begin{aligned} \alpha_j &= -\frac{\mu_j^2}{2\sigma_j^2} + \frac{\mu_1^2}{2\sigma_1^2} + \log\frac{\pi_j}{\pi_1} - \log\frac{\sigma_j}{\sigma_1} \\ \beta_j &= \frac{\mu_j}{\sigma_j^2} - \frac{\mu_1}{\sigma_1^2} \\ \gamma_j &= -\frac{1}{2\sigma_j^2} + \frac{1}{2\sigma_1^2} \end{aligned}$$

f) Hvilke antagelser må en forutsette for å komme tilbake til en modell uten kvadratiske ledd?

Tabellene nedenfor viser resultatene fra en logistisk modell-tilpasning.

Tabell 3: $\{(x_{i,1}, y_i), y_i \in \{1, 2\}\}$, x_1 som forklaringsvariabel

```
Coefficients:
                  Value   Std. Error    t value
(Intercept)  -2.57956529  0.681893723  -3.782943
         x1   0.01556107  0.003974543   3.915185

    Null Deviance:  95.52384 on 68 degrees of freedom
Residual Deviance:  62.60177 on 67 degrees of freedom
```

Tabell 4: $\{(x_i, y_i), y_i \in \{1, 2\}\}$, x_1, x_1^2 som forklaringsvariabel

```
Coefficients:
                    Value     Std. Error    t value
(Intercept)  -3.975288e+00  1.0266142648  -3.872232
         x1   3.092048e-02  0.0082101426   3.766132
       x1^2  -3.192519e-05  0.0000114153  -2.796703

    Null Deviance:  95.52384 on 68 degrees of freedom
Residual Deviance:  57.65831 on 66 degrees of freedom
```

Tabell 5: $\{(x_{i,1}, y_i), y_i \in \{1, 3\}\}$, x_1 som forklaringsvariabel

```
Coefficients:
                  Value   Std. Error    t value
(Intercept)  -1.22299237  0.553260993  -2.210516
         x1   0.01518804  0.004054723   3.745766

    Null Deviance:  133.6729 on 108 degrees of freedom
Residual Deviance:  114.1169 on 107 degrees of freedom
```

(Fortsettes side 6.)

Tabell 6: $\{(x_{i,1}, y_i), y_i \in \{1,3\}\}$, x_1, x_1^2 som forklaringsvariabel

```
Coefficients:
                     Value     Std. Error    t value
(Intercept)  -3.570726e+00  0.9311933588  -3.834570
         x1   4.721563e-02  0.0099266129   4.756469
       x1^2  -8.528434e-05  0.0000202593  -2.796703

    Null Deviance:  133.6729 on 108 degrees of freedom
Residual Deviance:  96.31143 on 106 degrees of freedom
Residual Deviance:  34.84639 on 106 degrees of freedom
```

Hvordan ville du gå frem for å teste om $\gamma_2 = \gamma_3 = 0$?
Hva blir resultatet av en slik test her?

g) Anta nå en ønsker å innføre restriksjonen

$$\sigma_2 = \sigma_3$$

mens σ_1 kan være vilkårlig.

Hvilke problemer får en når en gjør separate tilpasninger på datasettene

$$\{(x_{i,1}, y_i), Y_i \in \{1,2\}\}$$

og

$$\{(x_{i,1}, y_i), Y_i \in \{1,3\}\}?$$

Diskuter en alternativ metode der en kombinerer disse datasettene og gjør all estimering i ett.

Hva blir modell-likningen i dette tilfellet?

Oppgave 3.

Anta $Y_1, \cdots, Y_n$ u.i.f. med sannsynlighetstetthet

$$f(y) = \theta_2 e^{-\theta_2(y-\theta_1)}\, I(y > \theta_1)$$

Vi vil i første omgang se på testing av hypotesen

$$H_0 : \theta_1 = 0 \quad \text{mot} \quad H_A : \theta_1 > 0$$

a) Vis at problemet er invariant overfor transformasjonen

$$Y' = aY \quad \text{der} \quad a > 0$$

(Fortsettes side 7.)

b) Bruk suffisiens og invarians-betraktninger til å vise at testen bør avhenge av X_1/X_2 der

$$X_1 = \min(Y_1, \cdots, Y_n) \quad \text{og} \quad X_2 = \sum_{i=1}^{n} Y_i \,.$$

Hvorfor kan vi like gjerne se på

$$T = \frac{nX_1}{X_2 - nX_1} \,?$$

c) La $z_1 = nX_{(1)}$ og $z_2 = X_2 - nX_1$. Vis at

$$f(z_1, z_2) = \theta_2 e^{-z_1\theta_2} I(z_1 > 0) \theta_2^{n-1} e^{-z_2\theta_2} I(z_2 > 0)$$

Bruk dette til å vise at

- z_1 og z_2 er uavh
- $2\theta_2 z_1 \sim \chi^2_2$
- $2\theta_2 z_2 \sim \chi^2_{2(n-1)}$

d) Konstruer den overalt sterkeste invariante test basert på T.

Anta nå vi ønsker å teste

$$H_0' : \theta_2 = 1 \quad \text{mot} \quad H_A' : \theta_2 \neq 1$$

e) Vis at problemet er invariant under transformasjonen

$$Y' = Y + b$$

og at en maksimal invariant basert på de suffisiente observatorer er $X_2 - nX_1$.

Bruk dette til å konstruere den overalt sterkeste invariante test for H_0' mot H_A'.

SLUTT