UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Examination in:	STK2000 — Some central models and methods in statistics.
Day of examination:	Friday, December 10, 2004.
Examination hours:	09.00 – 12.00.

This examination set consists of 3 pages.

Appendices:	Table for the normal distribution, table for χ^2-distribution.
Permitted aids:	Formula list for ST100 and ST110, approved kalkulator.

Make sure that your copy of the examination set is complete before you start solving the problems.

Problem 1.

Consider an exponential family of distributions where the probability density can be written

$$f(y;\theta) = \exp[yb(\theta) + c(\theta) + d(y)] \tag{1}$$

where θ is a scalar parameter, and b, c, d are known functions.

a) Express $E(Y)$ by $c(\theta)$ and $b(\theta)$. Here Y is a random variable with probability density (1). Explain why $c(\theta) = -\log\{\int \exp[yb(\theta) + d(y)]dy\}$.

b) Explain why the probability density of a random variable with a Normal distribution with expectation μ and variance 1, can be written as (1). Explain what the functions b, c and d are in this case.

   What are the corresponding expressions for the probability function of a random variable which has a Poisson distribution with expectation λ?

(Continued on page 2.)

c) Let $Y_1, \ldots, Y_N$ be independent, identically distributed observation whose probability density can be written as (1). Show that the maximum likelihood estimator (MLE) for θ satisfies

$$\bar{Y} = -\frac{c'(\hat{\theta})}{b'(\hat{\theta})} \quad \text{where} \quad \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

Find expressions for the MLE when the observations are Normally and Poisson distributed as in b).

d) What is the asymptotic distribution of the MLE when N is large? Sketch the main arguments of how the asymptotic distribution is derived when the probability density of the observations can be written as (1).

Problem 2.

The $2\times2\times2$ table below shows the classification of 123 patients suffering from diabetes. They are classified according to whether they were dependent on injection of insulin (1 = dependent, 2 = not dependent), wherer any relatives also suffered from diabetes (1= relatives with diabetes, 2 = no relatives with diabetes) and to the age when they were diagnosed as diabetes patients (1 = under 45 years of age, 2 = 45 or more).

Denote the factors as "insulin", "diafam" and "alder" respectively.

		Relatives with diabetes			
		Yes		No	
		Age at diagnosis		Age at diagnosis	
		< 45	≥ 45	< 45	≥ 45
Dependent of injection of insulin	Yes	6	6	16	8
	No	1	36	2	48

Let the observations be y_{jkl} where j, k, l are the levels of insulin, diafam and alder.

a) Formulate a logistic regression model with

- response: dependency on injections of insulin
- covariates: Age at diagnosis (alder) and whether any relatives suffer from diabetes or not (diafam)

As link fuction you shall use a logit link. Explain what this means. Also explain what the linear predictor looks like when you use a "corner-point" parameterization with $k = l = 1$ as reference category.

(Continued on page 3.)

b) Below is a part of an analysis of deviance table. The column with degrees of freedom of the deviance has been taken away. Fill out what is missing. Use the analysis of deviance table to argue that the model "intercept + alder" is satisfactory.

Terms	Resid. Df	Resid. Dev	Test	Df	Deviance
1	3	50.03359			
alder	2	0.04667		?	49.98693
alder + diafam	1	0.03929	+diafam	?	0.00738
alder * diafam	0	0.00000	+alder:diafam	?	0.03929

c) The estimates of the parameters are

```
Coefficients:
                Value Std. Error   t value
(Intercept)  1.99243  0.6153449  3.237908
      alder -3.78419  0.6796930 -5.567498
```

Express the estimated model in terms of relevant odds ratios. Explain how the result should be interpreted. Here a "corner-point" parameterization is used with level 1 as reference category, so that the coefficient is the estimate for the parameter of the level "$\geq$ 45 år".

d) Compute 95% confidence intervals for the relevant odds ratios.

e) Explain what deviance residuals are and how they are used. Find the deviance residual in the model "intercept + alder" for the levels "age less than 45 years of age" and "no relatives with diabetes".

In the rest of the problem the three factors will be symmetrically treated, and the theme is log-linear models.

f) Explain how one can formulate a log-linear model to analyze the $2\times2\times2$ table. Explain the relationship between the parameters of the logistic regression model "intercept+alder+diafam" and the corresponding log-linear model. Explain how the likelihood of the logistic regression model "intercept + alder + diafam" can be maximized by maximization of the appropriate Poisson likelihood.

END