UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK3100/STK4100 — Introduction to generalized linear models.
Day of examination:	Friday 14 December 2018.
Examination hours:	09.00 – 13.00.
This problem set consists of 4 pages.	
Appendices:	Formulas in STK3100/4100.
Permitted aids:	Approved calculator and collection of formulas for STK1100/STK1110 and STK2120.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

We assume that $V \sim \text{bin}(n, \pi)$, i.e. binomially distributed with n trials and probability of success π, and let $Y = V/n$. Then the probability mass function (pmf) of Y takes the form

$$P(Y = y) = \binom{n}{ny} \pi^{ny} (1-\pi)^{n-ny} \tag{1}$$

for $y = 0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$.

a) Show that the distribution of Y is in the exponential dispersion family. That is, show that (1) can be written on the form

$$\exp\{[\theta y - b(\theta)]/a(\phi) + c(y, \phi)\}, \tag{2}$$

and determine θ, $b(\theta)$, $a(\phi)$ and $c(y, \phi)$.

b) Let μ denote the expected value of Y. Use the expressions for $b(\theta)$ and $a(\phi)$ to show that $\mu = \pi$ and determine $\text{var}(Y)$.

We then assume that $V_1, V_2, \ldots, V_N$ are independent with $V_i \sim \text{bin}(n_i, \pi_i)$, and let $Y_i = V_i/n_i$ for $i = 1, 2, \ldots, N$. We consider a generalized linear model (GLM) for $Y_1, Y_2, \ldots, Y_N$ with canonical link function $\log\{\pi_i/(1-\pi_i)\} = \eta_i$ and linear predictor $\eta_i = \beta_0 + \beta_1 x_i$. Here $x_1, \ldots, x_N$ are known covariate values.

c) Derive an expression for the log-likelihood function $L(\beta_0, \beta_1)$, and show that the maximum likelihood estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the solutions of the equations

$$\sum_{i=1}^{N} n_i(Y_i - \pi_i) = 0 \qquad \text{and} \qquad \sum_{i=1}^{N} n_i x_i (Y_i - \pi_i) = 0,$$

where $\pi_i = e^{\beta_0 + \beta_1 x_i}/(1 + e^{\beta_0 + \beta_1 x_i})$.

(Continued on page 2.)

d) By a general result for maximum likelihood estimation for GLMs, we know that $(\widehat{\beta}_0, \widehat{\beta}_1)^{\mathrm{T}}$ is approximately bivariate normally distributed with mean $(\beta_0, \beta_1)^{\mathrm{T}}$ and a covariance matrix that equals $\boldsymbol{\mathcal{J}}^{-1}$, where $\boldsymbol{\mathcal{J}}$ is the expected information matrix. Find an expression for $\boldsymbol{\mathcal{J}}$.

Problem 2

Titanic was a British passenger liner that sank in the Atlantic Ocean 15 April 1912, after colliding with an iceberg. There were about 1300 passengers and 900 crew aboard, and more than 1500 of them died. In this problem we will use data on 1046 passengers to investigate how the probability of surviving the disaster depends on the age and sex of the passengers and at which class they traveled. (Passengers with no information about age are omitted from the analysis.)

The data file `titanic` that is used in the analysis has one line for each of the 1046 passengers and the following variables in the four columns:

- `Sex`: Sex (1 = male; 2 = female)
- `Cage`: Centered age (age − 30)
- `Class`: Passenger class (1 = first class; 2 = second class; 3 = third class)
- `Survived`: Survived or died (0 = died; 1 = survived)

a) We first fit a model with only main effects, where `Sex` and `Class` are defined to be categorical covariates (factors), while `Cage` is a numeric covariate. The result of this model fit is given below. Describe the model that we have used. Give an interpretation of the estimated intercept and the estimate for (centered) age.

```
Call:
glm(Survived~Sex+Cage+Class, family=binomial, data=titanic)

Coefficients:
               Estimate    Std. Error    z value    Pr(>|z|)
(Intercept)    -0.007567      0.165527     -0.046       0.964
Sex2            2.497845      0.166037     15.044     < 2e-16
Cage           -0.034393      0.006331     -5.433    5.56e-08
Class2         -1.280571      0.225538     -5.678    1.36e-08
Class3         -2.289661      0.225802    -10.140     < 2e-16

    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  982.45  on 1041  degrees of freedom
```

b) On the next page is given the result for a model with interaction between sex and passenger class. Explain why this model is to be preferred to the one in question a. Describe which effects sex and passenger class have for the probability of surviving the disaster.

(Continued on page 3.)

```
Call:
glm(Survived~Sex+Cage+Class+Sex:Class, family=binomial, data=titanic)

Coefficients:
                  Estimate   Std. Error    z value     Pr(>|z|)
(Intercept)       -0.234083     0.186230     -1.257       0.209
Sex2               3.886388     0.492375      7.893    2.95e-15
Cage              -0.038401     0.006743     -5.695    1.23e-08
Class2            -1.600280     0.301987     -5.299    1.16e-07
Class3            -1.576159     0.252514     -6.242    4.32e-10
Sex2:Class2        0.070407     0.630978      0.112       0.911
Sex2:Class3       -2.488805     0.540041     -4.609    4.05e-06

    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  931.99  on 1039  degrees of freedom
```

c) Finally we fit models with more interactions. A summary of the fits of these models is given in the analysis of deviance table below. Some of the numbers in the table have been replaced by question marks. Fill in the correct numbers for the question marks, and explain how you arrive at these numbers. Which of the four models would you prefer? (Give the reasons for your answer.)

```
> anova(fit1,fit2,fit3,fit4,test="LRT")

Analysis of Deviance Table

Model 1: Survived~Sex+Cage+Class+Sex:Class
Model 2: Survived~Sex+Cage+Class+Sex:Class+Cage:Class
Model 3: Survived~Sex+Cage+Class+Sex:Class+Cage:Class+Sex:Cage
Model 4: Survived~Sex+Cage+Class+Sex:Class+Cage:Class+Sex:Cage+Sex:Cage:Class
     Resid. Df    Resid. Dev    Df    Deviance    Pr(>Chi)
1         1039        931.99
2         1037        922.17     ?        ?        0.00739
3           ?         917.84     1      4.3308     0.03743
4         1034           ?       2      1.8661     0.39335
```

Problem 3

Let $Y_1, \ldots, Y_n$ be independent and normally distributed with common variance σ^2 and

$$\mu_i = E(Y_i) = \beta_0 + \beta_1(x_i - \bar{x}); \qquad i = 1, 2, \ldots, n; \tag{3}$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. We introduce the vectors $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^{\mathrm{T}}$, and write (3) on vector/matrix form

$$\boldsymbol{\mu} = E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}.$$

Here $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ and $\boldsymbol{X} = [\mathbf{1}_n\,,\, \boldsymbol{x} - \bar{x}\mathbf{1}_n]$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ and $\mathbf{1}_n$ is a n-dimensional vector of 1's.

(Continued on page 4.)

a) Show that the projection matrix onto the model space $C(\mathbf{X})$ takes the form

$$\boldsymbol{P}_1 = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + M^{-1}(\boldsymbol{x} - \bar{x}\mathbf{1}_n)(\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}},$$

where $M = \sum_{i=1}^n (x_i - \bar{x})^2$.

b) Use the result in question a to show that the vector of fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_1\boldsymbol{Y}$ may be written

$$\widehat{\boldsymbol{\mu}} = \overline{Y}\mathbf{1}_n + \widehat{\beta}_1(\boldsymbol{x} - \bar{x}\mathbf{1}_n),$$

where $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $\widehat{\beta}_1 = M^{-1}\sum_{i=1}^n Y_i(x_i - \bar{x})$.

If $\beta_1 = 0$ we have the null model with $\mu_i = E(Y_i) = \beta_0$ for $i = 1, \ldots, n$. The projection matrix for the null model is known to be $\boldsymbol{P}_0 = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$. We then have the orthogonal decomposition

$$\boldsymbol{Y} = \boldsymbol{P}_0\boldsymbol{Y} + (\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}$$

with corresponding sum of squares decomposition

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}. \tag{4}$$

c) Show that

$$\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} = M\widehat{\beta}_1^2,$$

and

$$\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y} = \sum_{i=1}^n \left[Y_i - \overline{Y} - \widehat{\beta}_1(x_i - \overline{x})\right]^2.$$

d) Use Cochran's theorem to show that

$$M\widehat{\beta}_1^2/\sigma^2 \tag{5}$$

and

$$\sum_{i=1}^n \left[Y_i - \overline{Y} - \widehat{\beta}_1(x_i - \overline{x})\right]^2/\sigma^2 \tag{6}$$

are independent and (non-central) chi-squared distributed. Determine the degrees of freedom for the statistics (5) and (6), and show that the non-centrality parameter of (6) is 0. (One may show that the non-centrality parameter of (5) equals $M\beta_1^2/\sigma^2$, but you should not show this.)

e) Derive an F-statistic for testing the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative hypothesis $H_A : \beta_1 \neq 0$, and determine the distribution of the test statistic under H_0 and under H_A.

END