UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100/STK4100 — Introduction to generalized linear models
Day of examination:	Monday November 30th 2015
Examination hours:	14.30 – 18.30

This problem set consists of 4 pages.

Appendices:	Tables for normal-, t-, χ^2-distributions
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

In this problem you shall consider models where the response is considered as binomially distributed $Bin(n, \pi)$. Let $\mu = n\pi$.

a) Express this as a generalized linear model where the frequency distribution has the form

$$c(y, \phi) \exp(\frac{\theta y - a(\theta)}{\phi}).$$

Explain what is meant by a link function.

b) Assume $y_i, n_i, x_{i,1}, \ldots, x_{i,p+1}$, $i = 1, \ldots, n$ are n set of observations where y_i are are the responses and $x_{i,1}, \ldots, x_{i,p+1}$ are the covariates. The responses are assumed to be independent $Bin(n_i, \pi_i)$ distributed. Let $\hat{\mu}_i = n_i \hat{\pi}_i$ be the fitted values. What is the deviance from fitting this model? How is it expressed in this case? What is the main use of the deviance?

Problem 2

In this problem you shall consider data of survivals from a study of treatment for breast cancer. The response is the numbers that survived for three years. The covariates were the four factors

- app: appearance of tumor, two levels 1=malignant, 2=benign
- infl: inflammatory reaction, two levels 1= minimal, 2=moderate or severe
- age: age of patients, three levels 1= under 50, 2= 50-69, 3=70 or older
- country: hospital of treatment, three levels, 1= Japan, 2= US, 3= UK

(Continued on page 2.)

The number of survivors is modeled as a binomially distributed variable using a canonical logit link. Level 1 is used as base level or reference category for all factors.

a) The output from fitting the model where only appearance and country are used as covariates, i.e. a model with predictor of the form

$$\eta = \beta_0 + \beta_1 \text{fapp} + \beta_2 \text{fcountry2} + \beta_3 \text{fcountry3}$$

is displayed below. What is the interpretation of the estimate of the coefficient of appearance, `fapp` (`f` means factor)? Explain also how the coefficient can be expressed in terms of an odds ratio.

```
Call:
glm(formula = cbind(surv, nsurv) ~ fapp + fcountry, family = binomial,
    data = brc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8033  -0.7267   0.2157   0.7579   1.8742

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0811     0.1656   6.529 6.63e-11 ***
fapp2         0.5140     0.1659   3.098 0.001949 **
fcountry2    -0.6616     0.1993  -3.319 0.000902 ***
fcountry3    -0.4946     0.2071  -2.389 0.016917 *
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 57.637  on 35  degrees of freedom
Residual deviance: 36.662  on 32  degrees of freedom
```

b) The output below is an analysis of deviance table for comparing various model specifications. Fill out the positions indicated by a question mark.

(Continued on page 3.)

```
Analysis of Deviance Table

Model 1: cbind(surv, nsurv) ~ fapp + fage + fcountry
Model 2: cbind(surv, nsurv) ~ fapp + fage + finfl + fcountry
Model 3: cbind(surv, nsurv) ~ fapp + finfl + fage * fcountry
Model 4: cbind(surv, nsurv) ~ fapp * finfl + fage * fcountry
Model 5: cbind(surv, nsurv) ~ fapp * finfl + fapp * fage + fage * fcountry
Model 6: cbind(surv, nsurv) ~ fapp * finfl * fage * fcountry
  Resid. Df Resid. Dev Df Deviance
1        30     33.198
2         ?     33.197  1   0.0009
3        25     25.718  ?   7.4790
4        24     25.511  1        ?
5        22     22.059  2   3.4519
6         0      0.000  ?  22.0587
```

In the remaining parts of this problem we return to the model in part a) and consider the hypothesis

$$H_0 : \beta_2 + \beta_3 = -1 \text{ versus } H_a : \beta_2 + \beta_3 \neq -1.$$

c) The estimated covariance matrix between the estimators of the coefficients β_2 and β_3 is $\begin{pmatrix} 0.040 & 0.021 \\ 0.021 & 0.043 \end{pmatrix}$. Use a Wald test to test the null hypothesis above.

d) Explain how the null hypothesis can be tested with a likelihood ratio test by fitting two suitable models. No numerical calculations are necessary, but it must be specified how the pedictors should be defined.

Problem 3

The data used in this problem is for expenses in the the social security system Medicare in US. Average expenses per hospitalization, denoted as `ccpd`, were in six years recorded for 54 regions: the fifty states, Puerto Rico, Virgin Islands, District of Columbia and an unspecified other. Thus there are $6 \times 54 = 324$ observations. The expenses are treated as response. The covariates are `j = YEAR` which can take values $1, \ldots, 6$ and a factor indicating the average length of stay at hospital, `AVETD` in each region and year. This factor has tree levels, 1= six days or less, 2= 7-9 days, 3= 10 days or more. Six days or less is the base level and the others are denoted as AVETD_2 and AVETD_3.

Below the output from fitting the linear mixed effects model

$$y_{ij} = \beta_0 + \beta_1 \times j + \beta_2 \text{AVETD}_{2ij} + \beta_3 \text{AVETD}_{3ij} + b_{1i} + j \times b_{2i} + \varepsilon_{ij}, \quad j = 1, \ldots, 6 \, , i = 1, \ldots, 54$$

is displayed

(Continued on page 4.)

```
Linear mixed-effects model fit by REML
 Data: medicare
      AIC      BIC   logLik
  5200.98 5231.127 -2592.49

Random effects:
 Formula: ~1 + YEAR | fstate
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 2410.7972 (Intr)
YEAR         262.7191 0.418
Residual     429.6119

Fixed effects: ccpd ~ YEAR + factor(AVETD)
                   Value Std.Error  DF   t-value p-value
(Intercept)    7419.853  386.0518 267 19.219839  0.0000
YEAR            706.045   39.5543 267 17.849996  0.0000
factor(AVETD)2  567.721  183.9157 267  3.086857  0.0022
factor(AVETD)3 1008.339  244.2480 267  4.128342  0.0000
 Correlation:
               (Intr) YEAR   f(AVETD)2
YEAR            0.170
factor(AVETD)2 -0.488  0.168
factor(AVETD)3 -0.468  0.239  0.781

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.19288486 -0.60341726  0.01798739  0.61830554  3.51342658

Number of Observations: 324
Number of Groups: 54
```

a) Formulate the model in matrix form and explain what the usual assumptions are.

b) Compute a 95% confidence interval for the fixed effect coefficient for `YEAR`.

c) Explain how a test for simplifying the model by removing the random effect $b2$ can be performed.

d) What is the expectation and covariance matrix in the marginal model, i.e. of the response $(y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}, y_{i6})'$?

e) Explain how the null hypothesis $H_0 : \beta_3 = 2 \times \beta_2$ versus the alternative hypothesis $H_a : \beta_3 \neq 2 \times \beta_2$ can be tested? In this part no numerical calculations are expected.

END