UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100/STK4100 — Introduction to generalized linear models
Day of examination:	Wednesday November 30th 2016
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 gamma distributed, i.e. $G(\mu, \nu)$. The density function is

$$f(y; \mu, \nu) = \frac{y^{-1}}{\Gamma(\nu)}(\frac{y\nu}{\mu})^{\nu} \exp(-y\nu/\mu),\ y > 0.$$

a) Express this as an exponential distribution where the density function has the form

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

Identify $\theta, \phi, a(\theta)$ and $c(y, \phi)$.

b) Explain what the canonical link function is in this case, and discuss its use.

Problem 2

In the year 2014 147 persons were killed in road accidents in Norway. The figures classified according to gender and eight age groups can be found at the end of the problem set together with the size of the population in each group.

The output from fitting a model assuming that the responses were Poisson distributed

```
mod1<-glm(killed~factor(gender) + factor(age) +offset(log(population/100000),
                family=poisson,data=accidents)
```

is displayed below. The factors `gender` with male as base level and `age` group with 0-17 as base level, are used as covariates. The link function is the canonical link. Remark also that in the command the population size is divided by 100 000, so rates must be interpreted per 100 000 individuals.

(Continued on page 2.)

```
Call:
glm(formula = killed ~ factor(gender) + factor(age)
+ offset(log(population/1e+05)),family = poisson, data = accidents)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.09971  -0.40719  -0.00551   0.57037   1.05081

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.1506     0.3367   0.447 0.654634
factor(gender)2  -1.0212     0.1858  -5.495 3.91e-08 ***
factor(age)2      1.5565     0.4082   3.813 0.000137 ***
factor(age)3      1.0102     0.4216   2.396 0.016584 *
factor(age)4      0.8869     0.4272   2.076 0.037918 *
factor(age)5      1.5366     0.3867   3.973 7.09e-05 ***
factor(age)6      1.3329     0.4082   3.265 0.001095 **
factor(age)7      1.9813     0.3868   5.123 3.01e-07 ***
factor(age)8      2.1126     0.3988   5.297 1.18e-07 ***
---


(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 92.6417  on 15  degrees of freedom
Residual deviance:  9.9502  on  7  degrees of freedom
```

a) Discuss why a Poisson model is reasonable in this case. Using the information from the output, how would you judge the model fit?

b) Why is it sensible to include an `offset` in the linear predictor in this case?

c) What is the interpretation of the estimate of the intercept, β_0? What is the p-value of a test that the coefficient of the gender effect is equal to -1?

d) What is the fitted value and residual for women in the age group 45-54 years?

e) The total number of women who were killed was 40. The sum of the fitted values for women of all age group is also 40. Explain why.

Problem 3

The data in this problem is based on four measurements of a particular bone for 5 randomly selected boys. The measurements were taken at 8, 8.5, 9 and 9.5 years.

The variables included are

- bone: length of the bone in millimeters

(Continued on page 3.)

- redage: age -8.75, i.e. age centered

Below is an excerpt from the output from fitting a linear mixed model (LMM) with the procedure `lme` in R where the length of the bone, `bone` is the response,

$$\texttt{bone}_{ij} = \beta_0 + \beta_1 \texttt{redage}_{ij} + b_{i,1} + \texttt{redage}_{ij} b_{i,2} + \varepsilon_{ij}, i = 1, \ldots, 5,\ j = 1, 2, 3, 4$$

where $\mathbf{b}_i = (b_{i,1}, b_{i,2})^T,\ i = 1, \ldots, 5$ represent the random effects.

```
Linear mixed-effects model fit by REML
 Data: bonedat
       AIC      BIC    logLik
  44.97205 50.31428 -16.48603

Random effects:
 Formula: ~1 + redage | boy
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 0.8172867 (Intr)
redage      0.7323611 0.586
Residual    0.2939400

Fixed effects: bone ~ 1 + redage
             Value Std.Error
(Intercept) 52.690 0.3713644
redage       1.424 0.3479866
 Correlation:
       (Intr)
redage 0.543

Number of Observations: 20
Number of Groups: 5
```

a) Formulate the model on matrix form and explain the meaning and interpretation of the different parts. State the usual model assumptions carefully.

b) Use the values in the R-output to describe the distribution of the response $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})^T,\ i = 1, \ldots, 5$.

c) Find the conditional expectation of a random effect $\mathbf{b}_i,\ i = 1, \ldots, 5$ given the observations, i.e. $E[\mathbf{b}_i | \mathbf{y}_1, \ldots, \mathbf{y}_5]$. Describe how the random effects, $\mathbf{b}_i,\ i = 1, \ldots, 5$, can be predicted/estimated.

In part b) and c) it is not necessary to do any numerical calculations.

(Continued on page 4.)

Observed number of killed in road traffic in 2014 and size of Norwegian population according to gender, 1= male, 2= female, and age 1=0-17 , 2=18-24, 3=25-34, 4=35-44, 5=45-54, 6=55-64, 7=65-74, 8=75+

	population	killed	gender	age
1	576584	5	1	1
2	243510	11	1	2
3	349669	10	1	3
4	370541	11	1	4
5	359178	24	1	5
6	301981	12	1	6
7	224471	19	1	7
8	141500	15	1	8
9	548577	4	2	1
10	230407	7	2	2
11	333730	5	2	3
12	348593	3	2	4
13	338505	2	2	5
14	295223	6	2	6
15	232997	7	2	7
16	213610	6	2	8

END