UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100/STK4100 — Introduction to generalized linear models
Day of examination:	Monday December 1th 2014
Examination hours:	14.30 – 18.30
This problem set consists of 6 pages.	
Appendices:	None
Permitted aids:	Collection of formulas for STK1100/STK1110, STK2120 and approved calculator

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

Each subtask indexed by letters (1a, 1b etc.) counts equally. Each question numbered with Roman numerals (i), ii) and iii)) counts equally within each subtask.

Problem 1

1a

A distribution belongs to the exponential family if its probability mass function or probability density can be written in the form

$$f(y;\theta,\phi) = c(y,\phi)\exp[(\theta y - a(\theta))/\phi],$$

where $a(\cdot)$ and $c(\cdot,\cdot)$ are functions.

i) Show that if Y is a stochastic variable with a distribution belonging to the exponential family, then $\mathrm{E}(Y) = a'(\theta)$ and $\mathrm{Var}(Y) = \phi a''(\theta)$, where a' and a'' denote the first and second derivatives of a. [Hint: Start with calculating the first derivative of $f(y;\theta,\phi)$ with respect to θ.]

1b

The probability mass function for a Poisson distributed variable Y is

$$f(Y = y;\lambda) = (\lambda^y/y!)\exp(-\lambda).$$

i) Show that the Poisson distribution belongs to the exponential family.

ii) Show that $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda$.

(Continued on page 2.)

1c

Consider a regression problem with a Poisson distributed response variable Y, with logarithmic link function and with two explanatory variables x_1 and x_2 such that

$$Y \sim \text{Po}(\mu),$$
$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

i) Give an interpretation of the parameter β_1 or some transformation of it.

ii) Assume then that $\beta_0 = 1$, $\beta_1 = 2$ and $\beta_2 = 3$ and predict the response Y for $x_1 = 1$ and $x_2 = 1$ and then for $x_1 = 2$ and $x_2 = 1$.

1d

Consider now a specific data set with 100 observations of a count variable Y and two explanatory variables x_1 and x_2. The model in the previous exercise has been fitted to these data. Below you see some R output with information about the fitted model.

```
summary(glmobj)

Call:
glm(formula = y ~ x1 + x2, family = poisson(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-15.0942   -0.7773   -0.3345    0.5244   10.9833

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.863643   0.031640    27.3   <2e-16 ***
x1          2.132696   0.007939   268.6   <2e-16 ***
x2          2.970372   0.012227   242.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 264872.7  on 99  degrees of freedom
Residual deviance:   1063.8  on 97  degrees of freedom
AIC: 1372.6
Number of Fisher Scoring iterations: 4

> phihat<-sum(residuals(glmobj,type="pearson")^2)/(100-3)
> phihat
[1] 11.32317
```

i) Explain what over-dispersion means in Poisson regression.

ii) Explain why the results above show that the current count data are over-dispersed.

iii) Discuss shortly two different possibilities for performing a more correct analysis than that given above.

(Continued on page 3.)

Problem 2

2a

i) Give an interpretation of a regression coefficient β, or a transformation of it, in binary regression with logit link function.

ii) Give then a simpler interpretation of β which holds approximately for small probabilities.

2b

Consider a situation with 50 observations of a binary response variable Y and two continuous explanatory variables x_1 and x_2, where we fit models with different link functions and different explanatory variables included. Below you see the R code for fitting ten different models and the corresponding values of Akaike's Information Criterion (AIC).

```
> m0<-glm(y~1,family=binomial(link=log))
> m1.logit<-glm(y~x1,family=binomial(link=logit))
> m2.logit<-glm(y~x2,family=binomial(link=logit))
> m12.logit<-glm(y~x1+x2,family=binomial(link=logit))
> m1.probit<-glm(y~x1,family=binomial(link=probit))
> m2.probit<-glm(y~x2,family=binomial(link=probit))
> m12.probit<-glm(y~x1+x2,family=binomial(link=probit))
> m1.cloglog<-glm(y~x1,family=binomial(link=cloglog))
> m2.cloglog<-glm(y~x2,family=binomial(link=cloglog))
> m12.cloglog<-glm(y~x1+x2,family=binomial(link=cloglog))

> AIC(m0,
    m1.logit,m2.logit,m12.logit,
    m1.probit,m2.probit,m12.probit,
    m1.cloglog,m2.cloglog,m12.cloglog)

            df      AIC
m0           1 68.40641
m1.logit     2 55.38840
m2.logit     2 70.26156
m12.logit    3 57.11896
m1.probit    2 55.37494
m2.probit    2 70.26544
m12.probit   3 57.15451
m1.cloglog   2 55.62994
m2.cloglog   2 70.28034
m12.cloglog  3 57.56906
```

i) Define AIC.

ii) Which one of the models above would you choose based on the given results? Why?

2c

Assume that the data above are used to calibrate a test for diagnosing a disease, such that we predict that a patient has a disease ($Y = 1$) if the

(Continued on page 4.)

predicted probablity is larger than a threshold value γ.

i) Define the two terms sensitivity and specificity.

ii) Describe what a ROC (Receiver Operating Characteristics) curve is, and draw a plot with one curve for a model with good classification performance and another model which is no better than random classification.

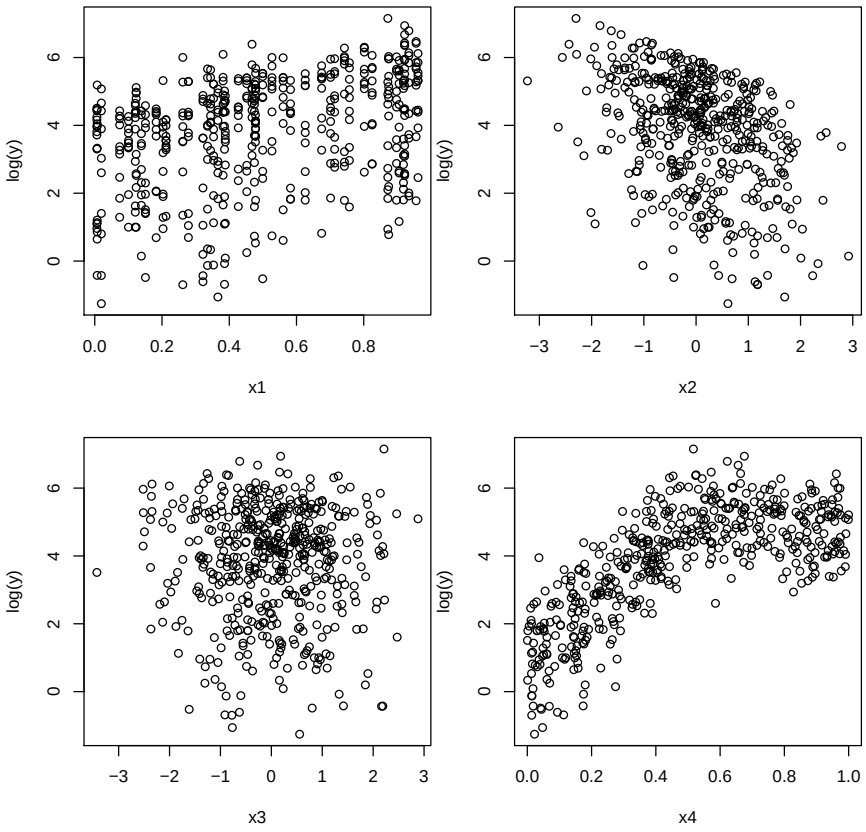
Problem 3

Consider a regression problem where

- The response Y is a continuous positive variable
- Y and corresponding explanatory variables are observed for 50 different groups with 10 observations within each group
- The continuous explanatory variable x_1 is group specific and has the same value within each group
- The three continuous explanatory variables x_2, x_3 and x_4 may have different values both between groups and between observations within the same group
- $\text{Var}(x_1) = 0.087$, $\text{Var}(x_2) = 0.98$, $\text{Var}(x_3) = 1.06$ and $\text{Var}(x_4) = 0.086$
- Groups are indexed by $i, i = 1, \ldots 50$ and observations within each group by $j, j = 1, \ldots 10$

(Continued on page 5.)

Below are scatter plots of the logarithm of the response vs. each of the explanatory variables.



We assume that the groups are a random subset of a population of groups. The following model has been fitted to these data

$$\begin{aligned} Y_{ij} &\sim \text{Gamma}(\mu_{ij}, \phi), \\ \log(\mu_{ij}) &= \beta_0 + b_i + \beta_1 x_{1i} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij}, \\ b_i &\sim N(0, \sigma_b^2), \end{aligned}$$

using the R code

```
> require(lme4)
> glmmobj<-glmer(y~x1+x2+x3+x4 + (1|g), family=Gamma(link=log))
```

Below you see a summary of the fitted object:

```
> summary(glmmobj)
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: Gamma  ( log )
Formula: y ~ x1 + x2 + x3 + x4 + (1 | g)

     AIC      BIC   logLik deviance df.resid
  4834.0   4863.5  -2410.0   4820.0      493
```

(Continued on page 6.)

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5533 -0.8675 -0.2161  0.7478  3.6526

Random effects:
 Groups   Name        Variance Std.Dev.
 g        (Intercept) 0.006691 0.0818
 Residual             0.353262 0.5944
Number of obs: 500, groups:  g, 50

Fixed effects:
            Estimate Std. Error t value Pr(>|z|)
(Intercept)  1.04721    0.08737   11.99   <2e-16 ***
x1           2.06204    0.10670   19.33   <2e-16 ***
x2          -0.64944    0.02861  -22.70   <2e-16 ***
x3           0.02241    0.02685    0.83    0.404
x4           4.14268    0.12616   32.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
   (Intr) x1     x2     x3
x1 -0.625
x2 -0.038  0.049
x3 -0.019  0.013 -0.049
x4 -0.710  0.032  0.008  0.004
```

3a

i) Discuss whether the random effect term b_i is an important part of the model compared to other parts of the model.

3b

i) Use the information you have to suggest simplifications or improvements of the model.

END