UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in	STK3100/4100 — Introduction to generalized l
Day of examination:	Thursday 6. desember 2012.
Examination hours:	14.30 – 18.30.
This problem set consists of 4 pages.	
Appendices:	Tabell over normal, χ^2 og t fordeling
Permitted aids: STK1100/STK1110 and STK2120	Accepted calculator. Formulae notes for

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

The different sub-points can mainly be solved independently. If you get stuck at one point, go further to the next point.

Problem 1

A stochastic variable Y follows a distribution in the exponential distribution family if the density (or the probability) for Y can be written as

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

For further calculations you can use that if

$$M_Y(t) = \mathrm{E}[\exp(Yt)] = \int \exp(yt) f(y) dy$$

exist for all t in a neighborhood of 0, then

$$\mathrm{E}[Y^r] = M_Y^{(r)}(0)$$

where $M_Y^{(r)}(\cdot)$ is the r-th derivative of $M_Y(t)$ wrt t.

(a) Calculate the expectation and variance in the exponential distribution class.

We will in the rest of this exercise look at the inverse Gaussian distribution, given by

$$f(y) = \frac{1}{\sqrt{2\pi y^3\sigma}} \exp\left\{-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right\}, \quad y > 0$$

(Continued on page 2.)

(b) Show that this distribution belongs to the exponential family and show that $\theta = -1/(2\mu^2)$ and $a(\theta) = -\sqrt{-2\theta}$. Also identify ϕ and $c(y;\phi)$.

(c) Find the expectation and the variance in the inverse Gaussian distribution. Use this to discuss what situations such a distribution can be useful to apply.

What kind of constraints are there in the parameters involved?

(d) Assume now $Y_1, ..., Y_n$ are independent variables from a generalized linear model (GLM) with the inverse Gaussian distribution as response distribution. Explain what this means.

Explain in general what deviance means and discuss what deviance can be used to in a GLM setting.

(e) Explain what we mean by canonical link and what kind of advantages there is in using such link functions.

What is the canonical link for the inverse Gaussian distribution?

Problem 2

We will in this exercise look at the models of the following type:

$$\begin{aligned} Y_{ij} =& \beta_0 + b_{0,i} + (\beta_1 + b_{1,i})x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \qquad (*) \\ \boldsymbol{b}_i =& (b_{0,i}, b_{1,i})^T \sim N(\mathbf{0}, \boldsymbol{D}) \end{aligned}$$

where $i \in \{1, ..., N\}$ is a group index while $j \in \{1, ..., n_i\}$ is an index for repeated measurements within group. We here assume that all b- and ε-variables are independent of each other.

(a) What is this model called?

Discuss the usefulness of such models.

(b) What is the *marginal* model for $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$?

What advantages do we have in the existence of an explicit expression of the marginal distribution for $\mathbf{Y}_i$ with respect to estimation?

(c) Discuss the main principles for REML estimation.

Discuss the advantages and disadvantages with maximum likelihood (ML) estimation compared to REML estimation. Specify in particular in which cases you would use the different methods.

Davidian and Giltinan (1995) describe a dataset for comparison of two types of soybeans. The dataset consists of 412 observations divided into 48 plots with 8-10 observations within each plot. In addition to weight (`weight`) and type soybean (`Variety`, two types) also the time the sample was taken (days after planting, `Time`) are given. The variable `Plot` specifies plot. In addition we define the variable `Time2` which is `Time` squared.

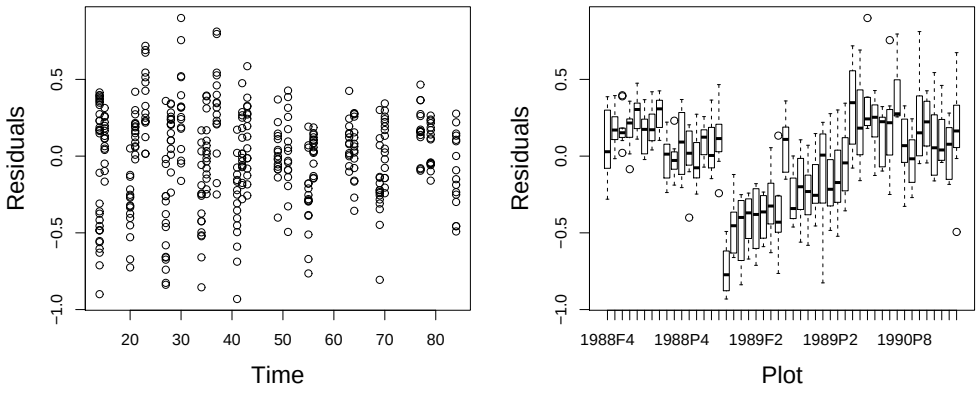
(Continued on page 3.)

(d) We will first look at a simple model where weight on log scale is used as response variable while `Variety`, `Time` and `Time2` are included as explanatory variables in addition to interaction between `Variety` and `Time`. We write the model in general as

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \varepsilon_{ij} \tag{M0}$$

where i specify plot while j specify repetition within plot.

The figure below shows boxplots of residuals grouped according to plots and a plot of residuals against `Time`. Comment on these plots and argue why a model similar to $(*)$ can be useful in this case.



(e) The model in the previous sub-exercise is now extended to two alternative models:

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i + \varepsilon_{ij} \tag{M1}$$

$$\log(\text{Weight}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_{0,i} + b_{1,i}\text{Time}_{ij} + \varepsilon_{ij} \tag{M2}$$

where $b_i \sim N(0, d^2)$ in model M1 and $\boldsymbol{b}_i = (b_{0,i}, b_{1,i})^T \sim N(\boldsymbol{0}, \boldsymbol{D})$ in model M2. The log-likelihood values (with REML estimates inserted) for the three models are given below:

Model	M0	M1	M2
Loglik	-130.92	-13.42	6.16

Based on this, argue why model M2 is preferable.

(f) Below are the results based on a fit of model M2 performed by ML estimation. Discuss if there is any need for simplifying this model.

```
Linear mixed-effects model fit by maximum likelihood

Random effects:
 Formula: ~1 + Time | Plot
             StdDev      Corr
(Intercept) 0.373119650 (Intr)
```

(Continued on page 4.)

```
Time        0.002970683 -0.999
Residual    0.190066092

Fixed effects: log(weight) ~ Variety + Time + Time2 + Variety:Time
                  Value  Std.Error  DF   t-value p-value
(Intercept)   -5.202444 0.09214869 361 -56.45707  0.0000
VarietyP       0.478989 0.11677518  46   4.10181  0.0002
Time           0.204265 0.00236086 361  86.52125  0.0000
Time2         -0.001317 0.00002330 361 -56.52636  0.0000
VarietyP:Time -0.003079 0.00124599 361  -2.47102  0.0139
```

END