UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK3100/STK4100 — Introduction to generalized linear models

Day of examination: Friday December 6th 2013

Examination hours: 14.30 – 18.30

This problem set consists of 3 pages.

Appendices: Table over normal distribution and table over χ^2-distribution

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.

Problem 1

a) Densities and frequency functions of the form

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

to describe the distribution of the response variable Y are used to define a generalized linear model (GLM). Describe the other parts of a GLM, and how they are related to $f(y;\theta,\phi)$.

b) Find a general expression for the expectation and the variance of the response variable Y.

c) What is meant with a saturated model? Explain how the deviance in a GLM is defined and explain how it can be used to compare two models. How is this procedure related to likelihood ratio tests? How are the deviance residuals defined?

d) Now consider the Poisson distribution and find expressions for the deviance and deviance residuals in this case.

Table 1: Observed numbers.

Number of children	0	1	2	3	4	5	6	Total
7 and 10 years educ.	128	161	194	61	12	6	2	564
13, 16 and 19 years educ.	44	26	37	5	2	1	0	115
All women	172	187	231	66	14	7	2	679

In the table above the number of children of 679 German women is recorded. The total number of children is 950. To investigate if there is any relation between years of education and number of children the material was divided into two parts: one with women with 7 and 10 years of education and

(Continued on page 2.)

another with those having 13, 16 or 19 years of education. Then two models were fitted using a GLM with Poisson distributed response and the canonical `log` link : one without any covariates, so the predictor was the same for all women, and another where a factor with two levels, that indicated to which groups the woman belonged, was the covariate. The number of children for each woman was considered as the response. The fitted values were 1.399 for the first model. For the second model the fitted values were 1.457 for those with 7 or 10 years education and 1.113 for the other.

e) Explain how the deviance can be used to test if there is any difference between the two groups with respect to the number of children. Carry out the test. Use a 5% level.

f) If $y_i,\ i = 1, \ldots, 679$ are the observed values and $\hat{\mu}_i,\ i = 1, \ldots, 679$ are the fitted values, explain why $\sum_{i=1}^{679}(y_i - \hat{\mu}_i) = 0$ for both models. Is this result also true for more general models?

Problem 2

The data in this problem is part of a longitudinal study of income in the US, the Panel Study of Income Dynamics, begun in 1968. The subset consists of 42 heads of household who were aged 25-39 in 1968. The variables included are

- annual nominal income, which is the response variable
- age, age in 1968
- cyear, coded as -10 in 1968,0 in 1978 and 10 in 1988
- educ, years of education in 1968
- sex, M=male, F=female

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

$$\texttt{lincm}_{ij} = \beta_0 + \beta_1 \texttt{age}_i + \beta_2 \texttt{cyear}_{ij} + \beta_3 \texttt{educ}_i + \beta_4 \texttt{sex}_i + b_i + \varepsilon_{ij}, i = 1, \ldots, 42,\ j = 1968, 1978, 1988$$

where b_i represent the random effects.

```
Linear mixed-effects model fit by REML
 Data: psid2
       AIC      BIC    logLik
  320.0178 339.5883 -153.0089

Random effects:
 Formula: ~1 | fid
        (Intercept)  Residual
StdDev:   0.0419192 0.7505293
```

(Continued on page 3.)

```
Fixed effects: lincm ~ age + cyear + educ + factor(sex)
                 Value Std.Error  t-value
(Intercept)   7.386823 0.6317104  11.693370
age          -0.020930 0.0152069  -1.376316
cyear         0.084163 0.0081889  10.277583
educ          0.116343 0.0275823   4.218021
factor(sex)M  1.311661 0.1422471   9.221007

Correlation:
             (Intr) age    cyear  educ
age          -0.831
cyear         0.000  0.000
educ         -0.685  0.201  0.000
factor(sex)M  0.003 -0.217  0.000  0.041

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-3.4411400 -0.4003431  0.1070887  0.5602338  1.6037724

Number of Observations: 126
Number of Groups: 42
```

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

b) Determine an approximate 95% interval for the coefficient of `cyear`. Do you think the nominal income has been constant in the period covered by the survey?

c) If one is interested in the simultaneous significance of two fixed effects, `age` and `educ` say, describe how that can be tested in this kind, LMM, of models.

d) Describe how the random effects b_i can be predicted/estimated?

e) Use the values in the R-output to calculate the estimated covariance matrix for the response $(Y_{i1}, Y_{i2}, Y_{i3})^T$.

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

END