UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK3100 / STK4100 — Introduction to generalized linear models.
Day of examination:	Wednesday December 2nd 2020
Examination hours:	15.00 – 19.30 (including 30 minutes for Inspera delivery).

This problem set consists of 4 pages.

Appendices:	None
Permitted aids:	All resources

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

Problem 1

a) Assume that Y is Poisson-distributed with $\mathrm{P}(Y = y; \lambda) = \frac{\lambda^y}{y!}\exp(-\lambda)$ for $y = 0, 1, \ldots$

Show that the Poisson-distribution can be written on the natural exponential family form $\mathrm{P}(Y = y; \lambda) = \exp(\theta y - b(\theta) + c(y))$. In particular identify the canonical parameter θ, the cumulant function $b(\theta)$ and $c(y)$.

Use $b(\theta)$ to derive expressions for $\mu = \mathrm{E}[Y]$ and $\mathrm{var}[Y]$.

b) Show that the truncated Poisson distribution for $Y \mid Y > 0$ has probability mass function $\frac{\lambda^y}{y!}\exp(-\lambda)/(1 - \exp(-\lambda))$ for $y = 1, 2, \ldots$.

Verify that also this distribution can be written on the natural exponential family form $\exp(\theta y - b(\theta) + c(y))$ and identify θ, $b(\theta)$ and $c(y)$ in this case.

c) As a generalization of this assume that Y has density or probability mass function $f(y; \gamma) = \exp(\gamma y - b_0(\gamma) + c_0(y))$ over a set S of permissible values of y and assume that B is a subset of S with $\mathrm{P}(Y \in B) > 0$ for all possible γ.

Show that in general $Y \mid Y \in B$ has a distribution on the natural exponential family form with density or probability mass function $f_B(y; \theta) = \exp(\theta y - b(\theta) + c(y))$ for $y \in B$. In particular identify θ and give an expression for $b(\theta)$.

(Continued on page 2.)

Problem 2

The data for this problem stems from an investigation of whether a health reform in Germany in 1997 led to reduced number of doctoral visits among women aged 20-40 years. Some individuals were interviewed in 1996 (before the reform) and others in 1998 (after the reform). The women reported the number of doctoral visits the last year. In this problem this response has been dichotomized into unfrequent and frequent visitors (somewhat arbitrarily) defined as $Y_i = 0$ if the number of visits were below 7 and $Y = 1$ if the number of visits were 7 or more. These responses Y_i are analysed with logistic regression for x_{i1} indicating interview before or after the reform (0=before, 1=after), x_{i2} = indicator of poor or very poor health (1=yes, 0=no, loosely referred to as "bad health") and other explanatory variables.

a) In a first model only x_{i1} and x_{i2}, denoted as `reform` and `badh` in the `R`-output below, were included in the logistic regression with binary outcomes defined in `R` as `I(numvisit>6)`.

   Give interpretations of $\exp(\hat{\beta}_j)$, where $\hat{\beta}_j$ are the estimates of the regression coefficient for x_{ij} for $j = 1$ and 2,

   (i) in general

   (ii) as an approximation valid when $\mathrm{P}(Y_i = 1)$ are all small (for these data the overall proportion $Y_i = 1$ was 0.086).

```
> fit=glm(I(numvisit>6)~badh+reform,family=binomial,data=drv)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7180     0.1230 -22.103   <2e-16
badh          2.1995     0.1668  13.189   <2e-16
reform       -0.3382     0.1619  -2.089   0.0367
```

b) Explain why $\exp(\hat{\beta}_j \pm 1.96 se_j)$, where se_j are the standard errors of the $\hat{\beta}_j$, are approximate 95% confidence intervals for $\exp(\beta_j)$.

   Calculate these confidence intervals for $\exp(\beta_1)$ and $\exp(\beta_2)$ and use the intervals to determine conclusions to hypothesis tests for $\mathrm{H}_{0j} : \beta_j = 0$ versus $\mathrm{H}_{0j} : \beta_j \neq 0$ with a 5% significance level.

   Compare the conclusions of the tests with the Wald z-values and p-values from the `R`-output.

(Continued on page 3.)

c) Below you find R-output from an extended model where also the explanatory variables education (categorized to three levels <10, 10-12 and > 12 years, `educat`) and income (categorized to three levels, `inccat`) as well as interactions between "bad" health and reform, "bad" health and income and education and income are included. The output is a deviance table with certain values replaced with question marks.

Explain what deviances are and what deviance table are used for.

Fill in correct values where 4 numbers are replaced by question marks.

```
> fit=glm(I(numvisit>6)~badh+reform+educat+inccat+badh:reform
       +badh:inccat+educat:inccat ,family=binomial,data=drv)
> anova(fit,test="Chi")

Analysis of Deviance Table
Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          2226     1303.4
badh          1  158.613      2225     1144.8 < 2.2e-16
reform        1        ?      2224     1140.4  0.035849
educat        2    2.339       ?          ?    0.310536
inccat        2    8.641      2220     1129.4  0.013292
badh:reform   1    1.458      2219     1127.9  0.227285
badh:inccat   2    0.851      2217     1127.1  0.653313
educat:inccat ?   13.689      2213     1113.4  0.008357
```

Problem 3

The log-normal distributions are not an exponential dispersion family. We will in this problem see how we may still use results or theory for GLM to fit regression models for responses that are log-normal.

a) Assume that Y is a log-normal random variable and so is defined by that $V = \log(Y)$ is a normal distributed random variable with mean $\mathrm{E}[V] = \gamma$ and $\mathrm{var}[V] = \sigma^2$.

Show that with $\mu = \mathrm{E}[Y]$ one can express $\mathrm{var}[Y] = \phi\mu^2$ where the dispersion parameter $\phi = \exp(\sigma^2) - 1$.

Hint: You can use that the moment-generating function of $V \sim \mathrm{N}(\gamma, \sigma^2)$ can be expressed as $M_V(t) = \mathrm{E}[\exp(tV)] = \exp(\gamma t + \frac{1}{2}t^2\sigma^2)$.

(Continued on page 4.)

b) Assume $Y_1, \ldots, Y_n$ are independent and log-normally distributed with expected values $\mu_i = \mathrm{E}[Y_i] = \exp(\alpha + \beta x_i + \frac{1}{2}\sigma^2)$ where the x_i are known explanatory variables with the same variance σ^2 of $V_i = \log(Y_i)$.

Demonstrate how simple linear regression for the $V_i = \log(Y_i)$ can be used to obtain estimates of α, β and ϕ.

c) If on the other side $\mu_i = \alpha + \beta x_i$ such a simple linear regression technique can not be applied to solve the estimation problem.

However one can use a technique from GLM-theory based on the relationship $\mathrm{var}[Y_i] = \phi\nu^*(\mu_i)$. Then $\nu^*(\mu_i)$ is a function specifying how the variances depend on the expected values $\mu_i = \mathrm{E}[Y_i]$. For the log-normal distribution we have from question a) that $\nu^*(\mu_i) = \mu_i^2$.

Explain in general terms this approach. It can be useful to state appropriate score equations for the estimation.

Problem 4

A model for binary matched pair data (Y_{i1}, Y_{i2}) with explanatory variables x_{ij} and random intercept u_i is written as

$$\mathrm{logit}(\mathrm{P}(Y_{ij} = 1 | x_{ij}, u_i)) = \beta_0 + \beta_1 x_{ij} + u_i$$

where $\mathrm{logit}(\pi) = \log(\pi/(1-\pi))$.
We assume that the u_i has a density $f(u; \sigma^2)$, typically from a $\mathrm{N}(0, \sigma_u^2)$-distribution, and that conditionally on u_i we have Y_{i1} and Y_{i2} independent. We also assume that the u_i's are independent so that the pairs $(Y_{i1}, Y_{i2}); i = 1, \ldots, n$ are independent.

a) Present an expression for the marginal probability $\mathrm{P}(Y_{ij} = 1 | x_{ij})$.

Argue that for each pair, Y_{i1} and Y_{i2} are marginally dependent.

Set up an expression for the marginal likelihood $l(\beta_0, \beta_1, \sigma_u^2)$.

b) Alternatively we may consider the u_i's as fixed effects and estimate β_1 by conditioning on $Y_{i1} + Y_{i2}$. Show that

$$\mathrm{P}(Y_{i1} = 1 | Y_{i1} + Y_{i2} = 1) = \frac{\exp(\beta_1(x_{i1} - x_{i2}))}{1 + \exp(\beta_1(x_{i1} - x_{i2}))}$$

and argue that $\mathrm{P}(Y_{i1} = 1 | Y_{i1} + Y_{i2} = 2) = 1 = \mathrm{P}(Y_{i1} = 0 | Y_{i1} + Y_{i2} = 0)$. (In these expression the conditioning on the observed numbers x_{ij} has been supressed from the notation).

Explain based on this how logistic regression can be set up to obtain an estimate of β_1.

END