Assignment 7, ST2304

Problem 1

The spreadsheet (http://www.math.ntnu.no/~diserud/ST2304/wildlife-encounters.csv) contains the following columns

t	How many years have you lived in Trondheim? Decimal numbers are ok.
area	In which part (center, east, west, south) of Trondheim have you mainly lived?
moose	Have you encountered any moose during your time in Trondheim?
fox	Any encounters with fox?
badger	Have you seen any badgers in Trondheim?
chanterelle	Any encounters with chanterell (kantarell) in the wild?
sex	Are you male of female?
studyprogram	biotech, biology or other
hours	How many hours per week spent in the wild if we define this as “utmark”?
nick	Optional nick name.

Fill in your own answers at the end of the spreadsheet. Use a value of 1 to indicate “yes” and 0 to indicate “no” for the variables moose, fox, and chanterelle. Note that all variables refer only to encounters within the municipality of Trondheim.

To analyze the data we will use the following model: We first assume that each individual encounters each of the species, say moose, according to a Poisson process with intensity λ (the encounter rate). The total number of encounters X during a time interval of length t is then Poisson distributed with parameter λt and the probability of at least one encounter is

$$p = P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{e^{-\lambda t}(\lambda t)^0}{0!} = 1 - e^{-\lambda t} \tag{1.1}$$

We also assume that the different explanatory variables have an additive effect on the log of the encounter rate λ, that is,

$$\ln(\lambda) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{1.2}$$

Equations (1.1) and (1.2) implies a model on the form

$$c \log\log(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \ln(t), \tag{1.3}$$

that is, a generalized linear model with a binomial response, a cloglog link-function and an offset term $\ln(t)$ (see handout 4 for details).

1. Fit models of this form in R using either moose, fox, badger or chanterelle as the response variable and area, sex, hours, and studyprogram as possible explanatory variables in each case. Remember to include `log(t)` as an offset term. Omit explanatory variables that are not

significant (use `drop1( ,test="Chisq")` to test this) and make sure that variables that are not included in the model are non-significant (use `add1( ,test="Chisq")`) to test this).

2. Based on the fitted model and the model assumption, what is the expected time you will need to wait until you encounter your next moose, fox, badger and chanterelle if you decide to stay in Trondheim?

Problem 2

Analyse how the variable menarche in the data set

```
juul.girl <- read.table("http://www.math.ntnu.no/~jarlet/statmod/menarche.dat")
```

depends on age using a generalized linear with a probit link function (see Dalgaard, p. 239 and handout 4 for details). This corresponds to the assumption that the underlying distribution of age T at which different individuals have their first period is normally distributed with a certain mean and standard deviations, say μ and σ .

1. Based on the estimate of the regression coefficients, compute the corresponding estimates of μ and σ .
2. Compute the variance and standard error of $\hat{\sigma}$ using the delta method (Handout 3).
3. Compute an estimate of the upper and lower 0.025-quantile of the distribution of T .

Problem 3

Load the data set

```
moose.ovulation <- read.table("http://www.math.ntnu.no/~jarlet/statmod/ovul2.dat")
```

into R. This data set (provided by Erling Solberg at NINA) consists of observations of moose shot during the hunting season in a local area of Norway. During the rut, female moose typically ovulates only once. In terms of management of moose populations, it is of interest to know when this occurs on average, as well as the degree of synchrony in time of ovulation. By examining whether or not a certain endocrine structure called the Corpus luteum is present, it can be determined whether or not ovulation has occurred at the time of death.

The dataset contains the following variables:

- time: The time of year each observation is made (number of days since January 1).
- n: The number of female moose shot at this particular day.
- x: The number out of these with corpus luteum present.

1. Compute the proportion x/n of individuals having ovulated at different days and plot this against the number of days since January 1. Since each observed proportion is based on different number of individuals, you may want to create a so called bubble plot by adjusting the size of each point in the graph using the additional argument `cex=sqrt(n)` in the call to `plot( )` so that each point covers an area proportional to n . You may want to adjust the size by multiplying `sqrt(n)` by a suitable scaling factor.

2. Fit a suitable generalized linear model to the data using time as the explanatory variable. Explain the rationale behind your choice of link function.
3. Add a curve representing the estimated relationship between time and the probability p to the plot. Also make a plot of the residuals of the model against time. Is there any systematic pattern in the residuals indicating that the model is wrong?
4. Assess the goodness-of-fit of the model based on the observed deviance (see handout 4). Can you reject the null hypothesis that the model is correct?
5. Estimate the proportion of female moose that should have ovulated by the end of the year according to the fitted model. Do you think this estimate is realistic?
6. If your conclusion is that the model is wrong, discuss briefly how it might be improved (difficult).