Solution of assignment 7, ST2304

Problem 1 1. Keeping only significant terms in the model using `drop1()` and `add1()` we end up with the following models using all 39 observations.

```
> summary(mod1)

Call:
glm(formula = moose ~ 1, family = binomial(link = "cloglog"),
    offset = log(t))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8617     0.2674  -6.962 3.36e-12 ***
---

> summary(mod2)

Call:
glm(formula = fox ~ area + hours, family = binomial(link = "cloglog"),
    offset = log(t))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.9378     1.2728  -3.879 0.000105 ***
areaeast      2.3813     1.2440   1.914 0.055584 .
areasouth     2.6166     1.3174   1.986 0.047013 *
areawest      2.5189     1.3417   1.877 0.060463 .
hours         0.2760     0.1225   2.254 0.024221 *

    Null deviance: 31.924  on 38  degrees of freedom
Residual deviance: 22.124  on 34  degrees of freedom
AIC: 32.124

> summary(mod4)

Call:
glm(formula = chanterelle ~ hours, family = binomial(link = "cloglog"),
    offset = log(t))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.0757     0.4213  -7.300 2.87e-13 ***
hours         0.2509     0.1051   2.386   0.0170 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 37.343  on 38  degrees of freedom
Residual deviance: 33.171  on 37  degrees of freedom
AIC: 37.171
```

```
> summary(mod4)

Call:
glm(formula = chanterelle ~ hours, family = binomial(link = "cloglog"),
    offset = log(t))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.0757     0.4213  -7.300 2.87e-13 ***
hours         0.2509     0.1051   2.386   0.0170 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 37.343  on 38  degrees of freedom
Residual deviance: 33.171  on 37  degrees of freedom
AIC: 37.171
```

2. Since we here have a Poisson process with parameter λ the expected waiting time T until the next encounter is exponential distributed with expectation

$$E(T) = \frac{1}{\lambda}. \tag{1}$$

According to the model

$$\lambda = e^{\eta} \tag{2}$$

where η is the linear predictor (not including the offset $\log t$).

The model for moose encounters only includes an intercept, hence $\hat{\lambda} = 1^{-1.86} = 0.156$ per year and $\hat{ET} = 1/\hat{\lambda} = 6.42$ years.

For fox encounters, the estimated value of the linear predictor η becomes

$$-4.93 + 2.38 + 0.27 \cdot 4 = -1.47 \tag{3}$$

for a person living in the Trondheim east, spending four hours in the wild each weak (Jarle). The corresponding value of λ and ET for such a person (Jarle) becomes $\hat{\lambda} = 1^{-1.47} = 0.230$ per year and $\hat{ET} = 1/\hat{\lambda} = 4.34$ years.

Similarly, the expected time ET until the next badger and chantarelle encounter becomes 1.97 and 7.92 years, respectively.

Problem 2 1. According to the model, the relationship betwen probability p of menarche having occured and age x is

$$\text{probit}\, p = \beta_0 + \beta_1 x. \tag{4}$$

From the model summary

```
> juul.girl <- read.table("http://www.math.ntnu.no/~jarlet/statmod/menarche.dat")
> summary(glm(menarche~age,fam=binomial("probit"),data=juul.girl))

Call:
glm(formula = menarche ~ age, family = binomial("probit"), data = juul.girl)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.37033    1.06346  -10.69   <2e-16 ***
age           0.86233    0.08106   10.64   <2e-16 ***
```

we have $\hat{\beta}_0 = -11.37$ and $\hat{\beta}_1 = 0.8623$. The mean and standard deviation of the underlying normal distribution latent age (see handout 4) of menarche are

$$\mu = -\beta_0/\beta_1, \qquad \sigma = 1/\beta_1 \tag{5}$$

so $\hat{\mu} = -13.19$ and $\hat{\sigma} = 1/.86 = 1.1596$ years.

2. To compute variance and standard error of $\hat{\sigma}$ we use the delta method. Here, the estimator $\hat{\sigma}$ is a function of $\hat{\beta}_1$ only,

$$\hat{\sigma} = f(\hat{\beta}_1) = \frac{1}{\hat{\beta}_1} \tag{6}$$

Thus,

$$\mathrm{Var}(\hat{\sigma}) \approx (\frac{\partial f}{\partial \beta_1})^2 \mathrm{Var}(\hat{\beta}_1). \tag{7}$$

The partial derivate of f with resepect to β_1 is

$$\frac{\partial f}{\partial \beta_1} = -\frac{1}{\beta_1}^2 = -\frac{1}{0.86^2} = -1.35 \tag{8}$$

Substituting this and the square of the standard error of $\hat{\beta}_1$ from the summary into into (7) we find that $\mathrm{Var}\,\hat{\sigma} = 0.012$ and $\mathrm{SE}(\hat{\sigma}) = 0.11$.

3. Since T is normally distributed, the upper and lower 0.025-quantile of the distribution can be found as follows

```
qnorm(c(0.025,0.975),13.19,1.16)
```

which gives the interval $(10.91, 15.46)$.

Problem 3 1. See Fig. 1.

2. The probit choice of link function corresponds to the assumption that time of ovulation of different female moose are normally distributed. Given that many different factors may affect the time of ovulation of each individuals (e.g. genetic factors, body condition, climatic condition etc.) we would expect the distribution to be approximately normal according to the central limit theorem. An alternative model based on the logit link function would imply a heavier tailed logistic distribution for the underlying time of ovulation which seems less realistic.

   The model summary for the model becomes

```
Call:
glm(formula = prop ~ time, family = binomial(link = "probit"),
    weights = n)
```

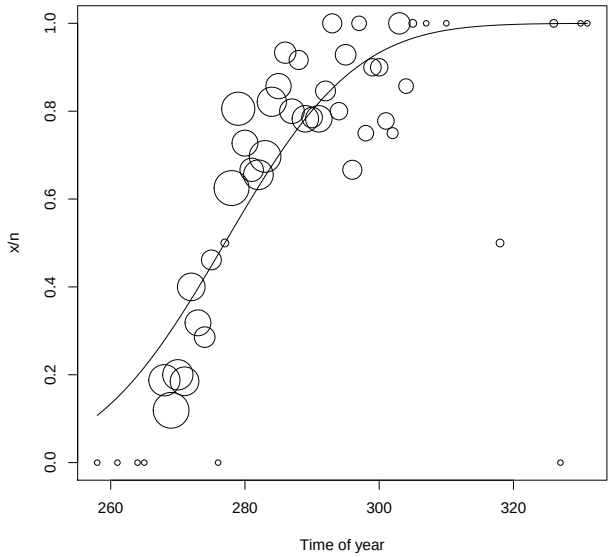

Figure 1: Proportion of x/n individuals having ovulated at different days against number of days since January 1, and the probability p

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.057365   1.642587  -10.99   <2e-16 ***
time          0.065188   0.005852   11.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 254.607  on 48  degrees of freedom
Residual deviance:  90.165  on 47  degrees of freedom
AIC: 189.29

Number of Fisher Scoring iterations: 6
```

3. Based on the model

$$\text{probit}\, p = \beta_0 + \beta_1 \text{time} \tag{9}$$

the p expressed as a function of time becomes

$$p = \phi(\beta_0 + beta_1 \text{time}) \tag{10}$$

where ϕ is the cumulative standard normal density function (denoted G in Løvås and `pnorm` in R).

The fitted model is shown in Fig. 1 and the residuals in Fig. 3. For certain time intervals the residuals are either almost always negative and positive indicating that the function relationship between p and time is wrong.

4. Given that the model is correct (H_0), the deviance of the model is chi-square with degrees of freedom equal to the residual degrees of freedom. Based on the large observed residual deviance of 90.16 we can reject this null hypothesis. The P value for the goodness-of-fit test becomes

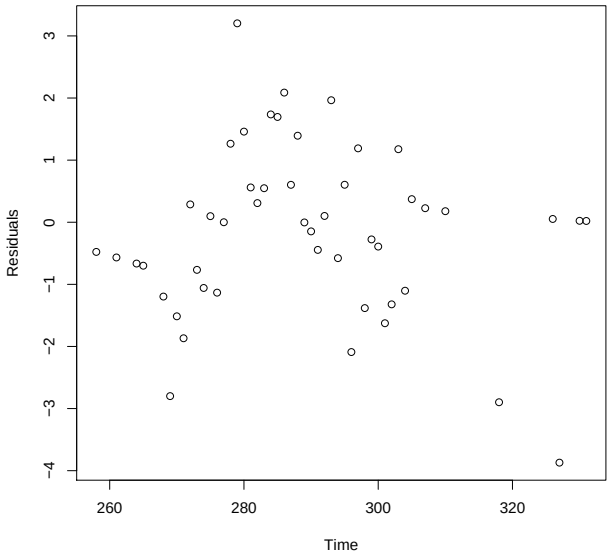

Figure 2: Residuals of model against *time*

```
> pchisq(90.165,47,lower=F)
[1] 0.0001551624
```

5. Using (10) we find that an estimate of p at the end of the year (time= 365) is

```
> pnorm(-18.05+0.065188*365)
[1] 1
> print(pnorm(-18.05+0.065188*365),digit=16)
[1] 0.9999999953663183
```

that is, very close to 1. This seems unrealistic based on the observed data (one individual had not ovulated on day 330).

6. The reason that the probit regression do not fit the data is that not all individuals ovulate such that p as a function of time get a sigmoid curve that flats out on a lower level then p=1. clearly the model is not very good, and maybe a logit-link function god give a better fit, assuming a logistic distribution with heavier tails than the normal distribution.

R code:

```
moose.ovulation <- read.table("http://www.math.ntnu.no/~jarlet/statmod/ovul2.dat")
attach(moose.ovulation)
prop=x/n
##make a plot
plot(time,prop,cex=sqrt(n)*0.8, xlab="Time of year", ylab="x/n")
##fit the model
mooselm=glm(prop~time, family=binomial(link="probit"), weight=n)
summary(mooselm)
##table of success and failures
mat<-cbind(x,n-x)
mooselm=glm(mat~time, family=binomial(link="probit"))
##add the curve of time and p
```

```
  curve(pnorm(-18.057365+ 0.065188 *x), to=max(time), from=min(time),add=T)
  ##plot the residuals
  plot(time,resid(mooselm), ylab="Residuals", xlab="Time")
##p value for the goodness-of-fit test
 pchisq(90.165,df=47,lower.tail=F)
 ##proportion ovualtion
  pnorm(5.736255)
```