Solution of assignment 7, ST2304

Problem 1 1. The model can be written as:

We choose chanterelle as response variable and fit the model as:

$$\begin{aligned}\text{cloglog}\, p_{\text{chanterelle}} = \beta_0 &+ \beta_{\text{area}}\text{area} + \beta_{\text{sex}}\text{sex} \\ &+ \beta_{\text{hours}}\text{hours} + \beta_{\text{studyprogram}}\text{studyprogram} + \ln \text{t}\end{aligned}$$

where area, sex and studyprogram are factors with levels

$$\text{area} = \text{center, east, south, west} \quad (1)$$
$$\text{sex} = \text{female, male} \quad (2)$$
$$\text{studyprogram} = \text{biology, biotech} \quad (3)$$
$$(4)$$

Not all these parameters can be estimated, therefore the constraint that the effect sizes in the control groups are zero is imposed, i.e. $\beta_{\text{areacenter}}$=0, $\beta_{\text{sexfemale}}$=0, and $\beta_{\text{studyprogrambiotech}}$.
The model can be rewritten as the multiple regression

$$\begin{aligned}\text{cloglog}\, p_{\text{chanterelle}} = \beta_0 &+ \beta_{\text{areaeast}}\ \text{areaeast}\ + \beta_{\text{areasouth}}\ \text{areasouth} \\ &+ \beta_{\text{areawest}}\text{areawest}\ + \beta_{\text{sexmale}}\text{sexmale} \\ &+ \beta_{\text{hours}}\text{hours} + \beta_{\text{studyprogrambiotech}}\text{studyprogrambiotech} + \ln \text{t}\end{aligned}$$

Use the `drop1` and `add1` function to omit explanatory variables non-significant variables. We end up with a model that have no significant variables only an intercept.

```
glmmchan5=glm(chanterelle~1,family=binomial(link="cloglog"),offset=log(t))
> summary(glmmchan5)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8803  -0.6806  -0.6648   0.5363   1.7071

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.4260     0.3083   -7.87 3.54e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 36.058  on 35  degrees of freedom
Residual deviance: 36.058  on 35  degrees of freedom
AIC: 38.058

Number of Fisher Scoring iterations: 4
```

Rcode:

```
data=read.table("wildlife-encounters.csv", header=T, sep=",")
attach(data)
glmmchan=glm(chanterelle~area+sex+studyprogram+hours,
         family=binomial(link="cloglog"), offset=log(t))

#checking the significance of variables
glmmchan=glm(chanterelle~area+sex+studyprogram+hours,
          family=binomial(link="cloglog"), offset=log(t))
drop1(glmmchan,test="Chisq")

glmmchan2=glm(chanterelle~area+studyprogram+hours,
          family=binomial(link="cloglog"), offset=log(t))
drop1(glmmchan2,test="Chisq")

glmmchan3=glm(chanterelle~studyprogram+hours,
         family=binomial(link="cloglog"), offset=log(t))
drop1(glmmchan3,test="Chisq")

glmmchan4=glm(chanterelle~hours,
        family=binomial(link="cloglog"), offset=log(t))
drop1(glmmchan4,test="Chisq")

glmmchan5=glm(chanterelle~1,
          family=binomial(link="cloglog"), offset=log(t))

##add1 function
glmmchanA=glm(chanterelle~1,
          family=binomial(link="cloglog"), offset=log(t))
add1(glmmchanA,.~.+area+sex+hours+studyprogram, test="Chisq")
```

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

$$E(\text{waiting time}) = \frac{1}{\lambda} \tag{5}$$

finding the encounter rate λ from equation (2) in assignment 7. Setting in for λ in our fitted model

$$\begin{aligned}\lambda &= e^{\beta_0} \\ &= e^{-2.4260} = 0.09.\end{aligned}$$

The expected waiting time to see a chanterelle is 32.85 days.

Problem 2 1. We have from handout 4 that

$$\text{probit } p = \beta_0 + \beta_{\text{age}} x_{\text{age}} \tag{6}$$

Having estimates for the regression estimates from summary() of the generalized model with probit link function we solve

$$\sigma = \frac{1}{\beta_{\text{age}}}, \mu = -\frac{\beta_0}{\beta_{\text{age}}} \tag{7}$$

and get $\mu = 13.19$ and $\sigma = 1.16$.

2. To compute variance and standard error of $\hat{\sigma}$ we use the delta method in handout 3, ligning (9).

$$\sigma = f(\beta_{\text{age}}) = \frac{1}{\beta_{\text{age}}} \tag{8}$$

Then the expression in equation (9) becomes

$$\text{Var}(\hat{\sigma}) \approx (\frac{\partial f}{\partial \beta_{\text{age}}})^2 \text{Var}(\hat{\beta}_{\text{age}}) \tag{9}$$

Finding the partial derivate of β_1

$$f'(\beta_{\text{age}}) = -\frac{1}{\beta^2_{\text{age}}} \tag{10}$$

Then the variance and standard error of σ becomes

$$\text{Var}(\hat{\sigma}) = (\frac{1}{\beta^4_{\text{age}}}) \text{Var}(\hat{\beta}_{\text{age}}) \tag{11}$$

$$\text{SE}(\hat{\sigma}) = (\frac{1}{\beta^2_{\text{age}}}) \text{SE}(\hat{\beta}_{\text{age}}) \tag{12}$$

We find the variance and standard error of β_1 in the summary(). The variance of $\hat{\sigma}$=0.012, and the standard error of $\hat{\sigma}$=0.11.

```
 > summary(period)
Call:
glm(formula = menarche ~ age, family = binomial(link = "probit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.32986  -0.15223   0.00028   0.07228   2.48281

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 719.39  on 518  degrees of freedom
Residual deviance: 197.39  on 517  degrees of freedom
AIC: 201.39

Number of Fisher Scoring iterations: 8
```

R code:

```
juul.girl <- read.table("http://www.math.ntnu.no/~jarlet/statmod/menarche.dat")
attach(juul.girl)
period=glm(menarche~age,family=binomial(link="probit"))
summary(period)
#variance
varsig=(1/(0.86233^4))*(0.08106^2)
#standard error
sdsig=(1/(0.86233^2))*(0.08106)
```

3. We calculate the upper and lower 0.025-quantile (95% confidence interval) of the distribution of T, using that T is normally distributed with known standard deviation of T.

   we use the quantile function in R for the normally distrbution: R code:

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

   giving the 0.025-quantile of T [10.91,15.46]

Problem 3 1.

2. We fit a generalized linear model with response variable x/n that have a binomial distribution and *time* as an explanatory variable, and weight the response with n. If we assume that the time of ovulations are normally distributed in the population we can use the probit link.

   Alternative: Make a table with successes and failures, and fit the model (without weights=n)

3.
```
> summary(mooselm)

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

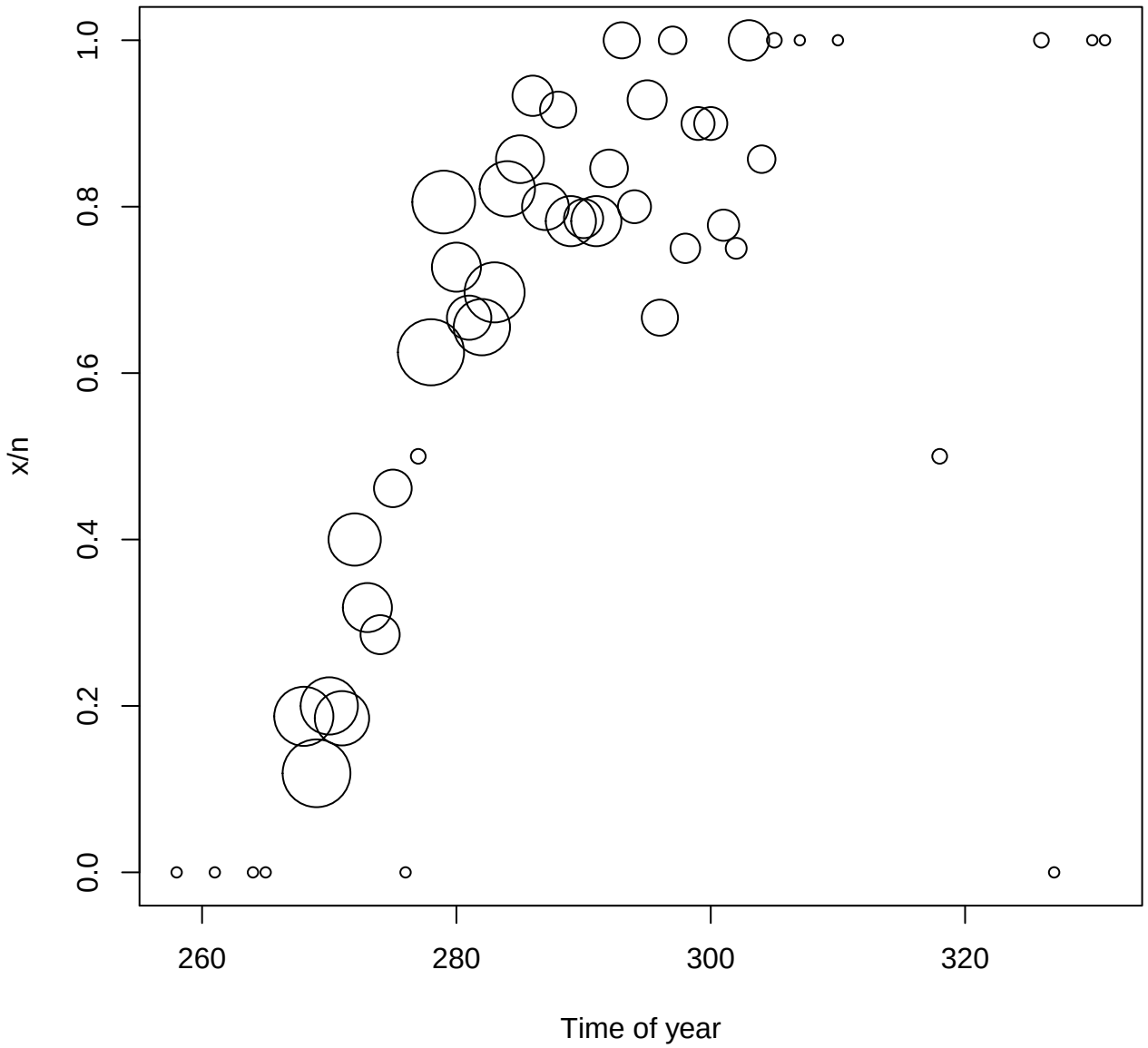

Figure 1: Proportion of x/n individuals having ovulated at different days against number of days since January 1.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8703  -1.0580   0.0004   0.5604   3.2028

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
```

4. The probability p

$$\text{probit}(p) = \beta_0 + \beta_{\text{time}}\text{time} \tag{13}$$

Setting in for the regression coefficient in `summary`() and obtaining an expression between time and p, we then add the curve to the plot.

The probit equation yields a area under the normal distribution of

$$\text{probit}(p) = -18.057365 + 0.065188\text{time} \tag{14}$$

We plot the residuals of the model against time, using the `resid` function in R. If seems like the residuals are positive on the interval from 280 to 295 days and almost entirely negative from 295 days, showing a trend in the residuals, something that suggest that the model do not fit the data.

5. H_0: fitted model is true (to the data). It follows that the deviance D is approximately chi-square distributed with $n - p$ (p regression coefficients) degrees of freedom. If D is sufficiently large we reject H_0 and conclude the model does not fit the data.

   >From the `Summmary`() we see that the deviance D is 90.165 (the residual deviance) with $n - p$=49-2=47 df.

   The p-value for the goodness-of-fit test becomes, 0.00016, and we can reject the null hypothesis. The model does not fit the data.

6. We set into the equation (9), for the regression coefficients and time=365 (days since January 1.).

$$\text{probit}(p) = -18.057365 + 0.065188 * 365, \tag{15}$$

and area under the normal distribution (probit link function is the inverse of the cumulative standard normal density), was found to be 1, every females moose have ovulated. This may seem strange, as not all individuals ovulate.

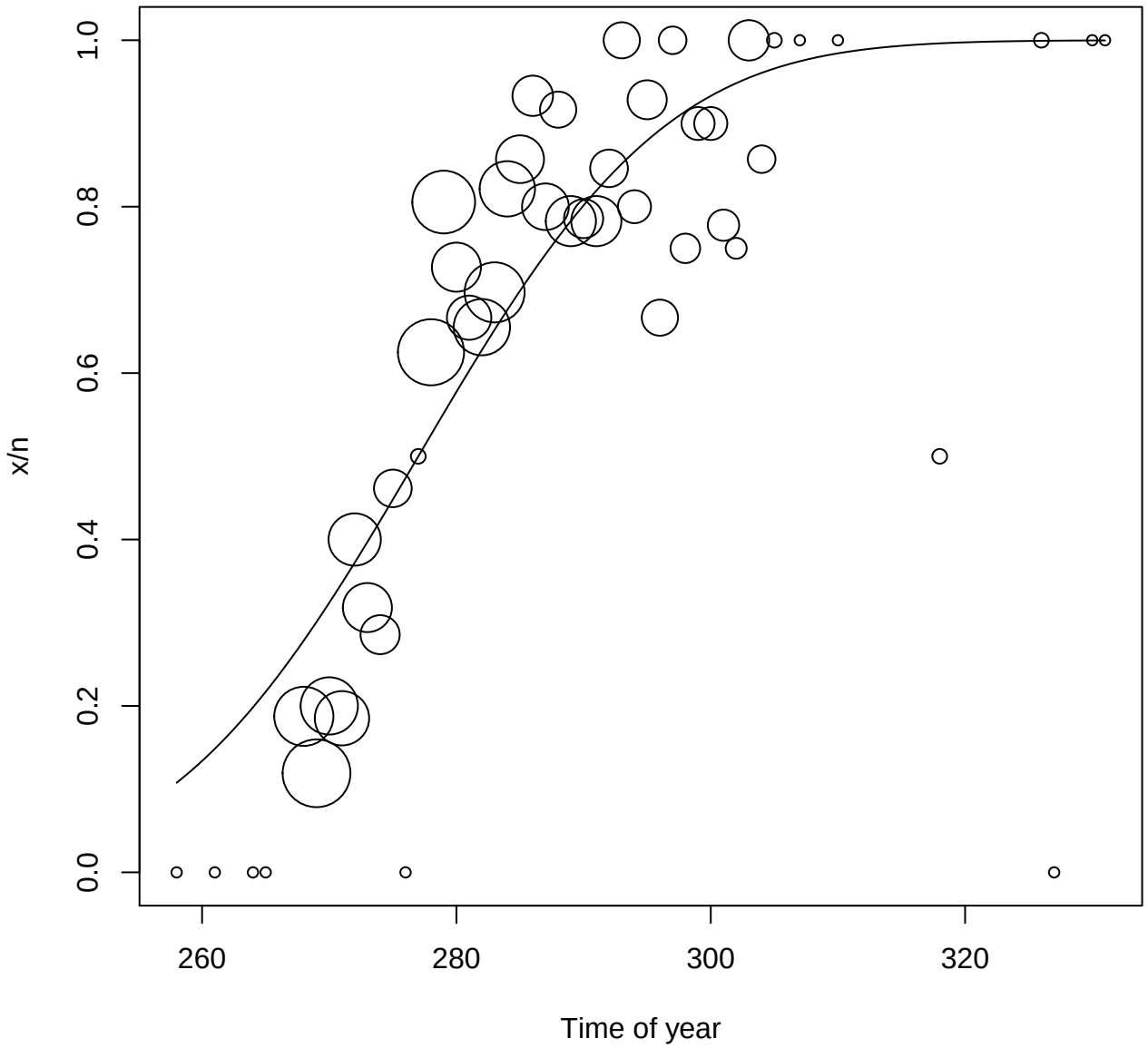

Figure 2: Proportion of x/n individuals having ovulated at different days against number of days since January 1, and the probability p

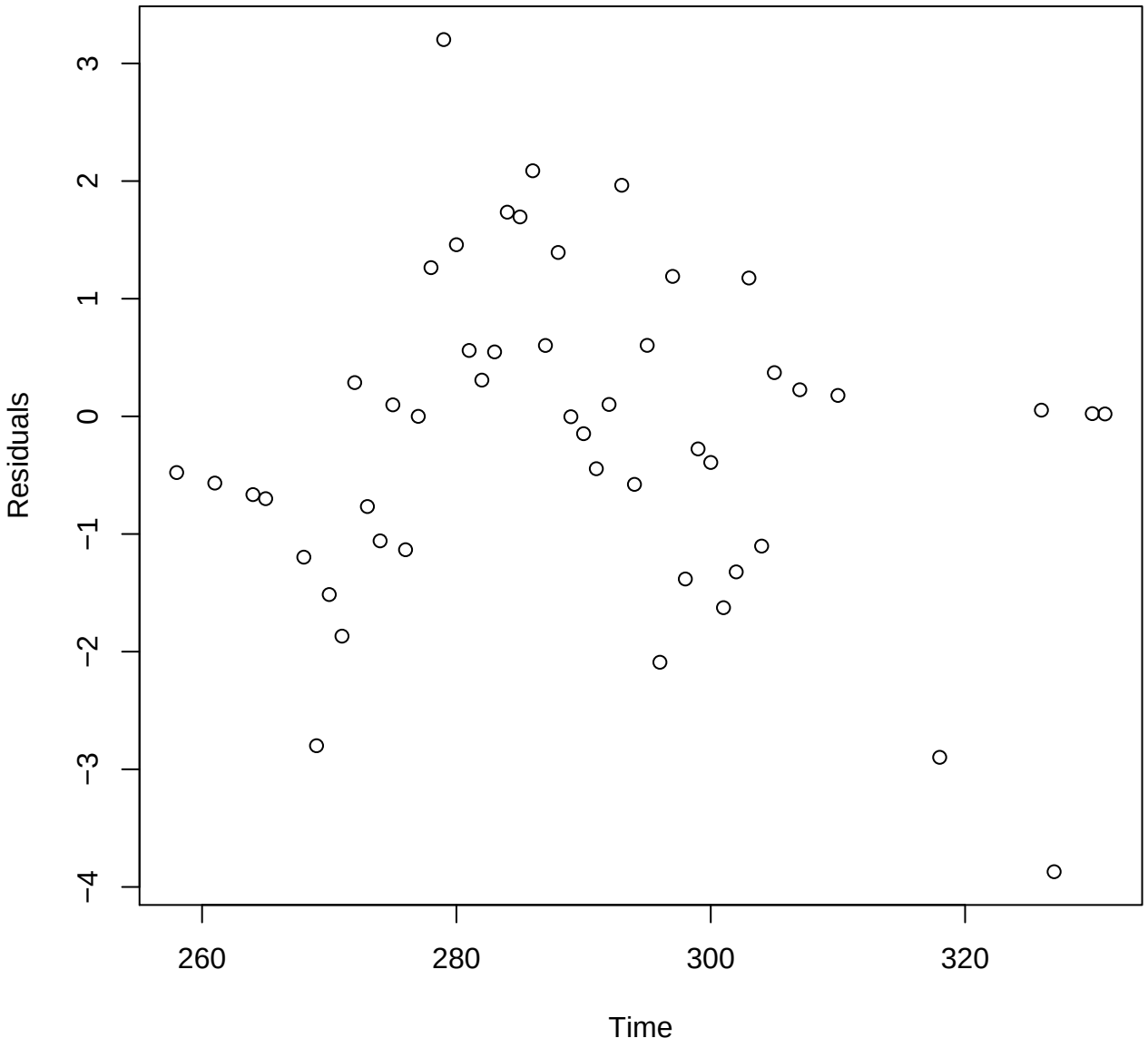

Figure 3: Residuals of model against *time*

7. 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)
```