Generalized Linear Models

Goals:
Learn how to construct and solve GLMs in R

Outline:
Simple glm (one independent variable)
Multiple glm (more than one independent variable)
Possibility for interactions

Assignments:

Horseshoe crab study

1a. Construct and solve glms with \textit{logit} and \textit{probit} links for the data from the following table:

<table>
<thead>
<tr>
<th>Female Crab Width, cm</th>
<th>Number of observations</th>
<th>Number having satellites</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>14</td>
<td>5</td>
<td>0.36</td>
</tr>
<tr>
<td>24</td>
<td>14</td>
<td>4</td>
<td>0.29</td>
</tr>
<tr>
<td>25</td>
<td>28</td>
<td>17</td>
<td>0.61</td>
</tr>
<tr>
<td>26</td>
<td>39</td>
<td>21</td>
<td>0.54</td>
</tr>
<tr>
<td>27</td>
<td>22</td>
<td>15</td>
<td>0.68</td>
</tr>
<tr>
<td>28</td>
<td>24</td>
<td>20</td>
<td>0.83</td>
</tr>
<tr>
<td>29</td>
<td>18</td>
<td>15</td>
<td>0.83</td>
</tr>
<tr>
<td>30</td>
<td>14</td>
<td>14</td>
<td>1.00</td>
</tr>
</tbody>
</table>

This table comes from a study of nesting horseshoe crabs (J. Brockmann, \textit{Ethology}, 102: 1-21, 1996). Each female horseshow crab in the study had a male crab attached to her in her nest. The study investigated whether the female crab width has any effect on whether the female had any other males, called satellites, residing nearby her.

1b. Plot the observed sample proportions and the predicted values from \textit{logit} and \textit{probit} models.

1c. Find the predicted number of satellites for the female width of 25.5 cm according to each model.

Snoring study (see below)

2. Use different numerical representation of snoring levels to construct \textit{logit} and \textit{probit} models for snoring data. Find such snoring levels that your models have the best fit to the data (by trial-and-error). Compare the best levels and discuss.

Pyrethroid study (see below)

3. Solve logistic models that use a) only sex, b) only dosage, c) linear combination of sex and dosage, but not their interactions, and, finally, d) sex, dosage, and interaction between sex and dosage. Compare the models, discuss, choose the model that you believe is the best for
describing the data and support your decision. (This *should not* be based on formal tests, like Akaike, *etc.*, use your feeling and R-graphs.)

**Reports:** Assignments require printed report, which will consist of R-results (do not print the entire session, only the necessary results!) and plots. Describe briefly the theoretical background for the methods you use, including necessary formulas, and make short statements about result interpretation. Consult instructor if you have any questions about the level of detail or formatting of your report.
**Essential R commands:**

**Session management:**
- `help()`
- `ls()`
- `getwd()`
- `setwd()`
- `library()`
- `data()`
- `save()`
- `load()`
- `read.table()`
- `class()`
- `names()`
- `rm()`

**Vectors:**
- `c()`
- `seq()`
- `rep()`
- `factor()`
- `cbind()`
- `rbind()`

**Data summaries:**
- `mean()`
- `sd()`
- `median()`
- `quantile()`
- `summary()`

**Graphs:**
- `par()`
- `plot()`
- `points()`
- `lines()`
- `mosaicplot()`
- `text()`

**GLMs:**
- `glm()`
- `family()`
- `summary.glm()`
- `predict.glm()`
1. **One independent variable**

We consider here an example based on the data from an epidemiological survey of 2484 subjects described in the table below:

<table>
<thead>
<tr>
<th>Snoring</th>
<th>Heart Disease</th>
<th>Proportion Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Never</td>
<td>24</td>
<td>1355</td>
</tr>
<tr>
<td>Occasional</td>
<td>35</td>
<td>603</td>
</tr>
<tr>
<td>Nearly every night</td>
<td>21</td>
<td>192</td>
</tr>
<tr>
<td>Every Night</td>
<td>30</td>
<td>224</td>
</tr>
</tbody>
</table>

The subjects here are classified according to their snoring level, as reported by their spouses (see P. G. Norton and E. V. Dunn, *Br. Med. J.*, 291, 630-632, 1985.)

Our goal is to construct a glm that will describe the above observations. We will consider logit and probit link functions that lead to the following models:

\[
\log\left( \frac{\pi(x)}{1-\pi(x)} \right) = \alpha + \beta x \quad \text{(logit)}
\]

\[
\text{probit}(\pi(x)) = \alpha + \beta x \quad \text{(probit)}
\]

For example, for snoring levels \(x=\{0, 3, 4.5, 5\}\) we have

\[
\log\left( \frac{\pi(x)}{1-\pi(x)} \right) = -4.05 + 0.38x
\]

\[
\text{probit}(\pi(x)) = -2.12 + 0.17x
\]

The sample code in program glm.R shows how to estimate parameters in those models and plot the results. The model fitting is done using the command glm(), the detailed results can be seen using the command summary(), the prediction using the fitted model is done using the command predict().

The details of using the package R will be discussed in the Lab.

2. **More than one independent variable**

In case when we have more than one independent variable (say, a and b), the analysis can take into account not only individual levels of those variables, but also their joint levels. In R, this is achieved by using expressions a:b or a:b. The sample code in program glm.R shows how to estimate parameters in those models and plot the results for the example on the toxicity to the tobacco budworm *Heliothis virescens* of doses of the pyrethroid trans-sypermethrin. The data are summarized in the table below:

<table>
<thead>
<tr>
<th>Sex</th>
<th>Dose, µg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>0</td>
</tr>
</tbody>
</table>
Here, batches of 20 moth of each sex were exposed for three days to the pyrethroid and the number in each batch that were dead or knocked down was recorded.
Sample R session (file glm.R)

#=========================================
#              STAT 453/653
#        Generalized Linear Models
#=========================================

# One independent variable
#=========================================

s<-c(0,3,4.5,5)
hd<-c(24,35,21,30)
n<-c(24+1355,35+603,21+192,30+224)
SF<-cbind(hd,n)
gl<-glm(SF~s,family=binomial(link='logit'))
gp<-glm(SF~s,family=binomial(link='probit'))
summary(gl)
summary(gp)

windows()
plot(c(0,5),c(0,.2),type='n',xlab='Snoring Level',ylab='Proportion')
points(s,hd/n,type='p',pch=19,col='red')
x<-seq(0,5,.1)
pl<-predict(gl,data.frame(s=x),type='response')
lines(x,pl,col=3)
pp<-predict(gp,data.frame(s=x),type='response')
lines(x,pp,col=4)

windows()
plot(c(0,5),c(-5,-1),type='n',xlab='Snoring Level',ylab='Logit')
points(s,log(hd/n/(1-hd/n)),type='p',pch=19,col='red')
x<-seq(0,5,.1)
pl<-predict(gl,data.frame(s=x),type='link')
lines(x,pl,col=3)

# Two independent variables
#=========================================

ldose<-rep(0:5,2)
numdead<-c(1,4,9,13,18,20,0,2,6,10,12,16)
sex<-factor(rep(c("M","F"),c(6,6)))
SF<-cbind(numdead,numalive=20-numdead)
g<-glm(SF~sex+ldose,family=binomial)
g<-glm(SF~sex*ldose,family=binomial)

windows()
plot(c(1,32),c(0,1),type='n',xlab='dose',ylab='proportion',log='x')
text(2^ldose,numdead/20,labels=as.character(sex))
ld<-seq(0,5,0.1)
lines(2^ld,predict(g,
data.frame(ldose=ld,
sex=as.factor(rep("M",length(ld)),levels=levels(sex))),
type='response'),col=3)
lines(2^ld,predict(g,
data.frame(ldose=ld),
sex = factor(rep("F", length(ld)), levels=levels(sex)),
type='response'), col=2)

windows()
plot(c(1, 32), c(-5, 5), type='n',
xlab='dose', ylab='proportion', log='x')
text(2^ldose, log(numdead/20/(1-numdead/20)), labels=as.character(sex))

ld <- seq(0, 5, 0.1)
p <- predict(g, data.frame(ldose=ld, sex=factor(rep("M", length(ld)), levels=levels(sex)),
type='response'))
lines(2^ld, log(p/(1-p)), col=3)
p <- predict(g, data.frame(ldose=ld, sex=factor(rep("F", length(ld)), levels=levels(sex)),
type='response'))
lines(2^ld, log(p/(1-p)), col=2)