_____ STAT 453/653, Fall 2012, Statistical Lab 3 _____

Generalized Linear Models: One explanatory variable

Goals:

Learn how to construct and solve GLMs with a single explanatory variable

Assignments:

Horseshoe crab study

1a. Construct and solve glms with *logit* and *probit* links for predicting the proportion of females having satellites from the female width using the following data:

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

This table comes from a study of nesting horseshoe crabs (J. Brockmann, *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 *logit* and *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 *logit* and *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.

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.

The deadline for reports is Monday, October 29, 2012

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:

Snoring	Heart Disease		Proportion Yes
	Yes	No	
Never	24	1355	0.017
Occasional	35	603	0.055
Nearly every night	21	192	0.099
Every Night	30	224	0.119

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 predict the probability to have a heart disease from the snoring level. We will consider *logit* and *probit* link functions that lead to the following models:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\alpha+\beta x \qquad \textbf{(logit)}$$

$$\text{probit}\left(\pi(\text{x})\right)=\alpha+\beta x \qquad \textbf{(probit)}$$

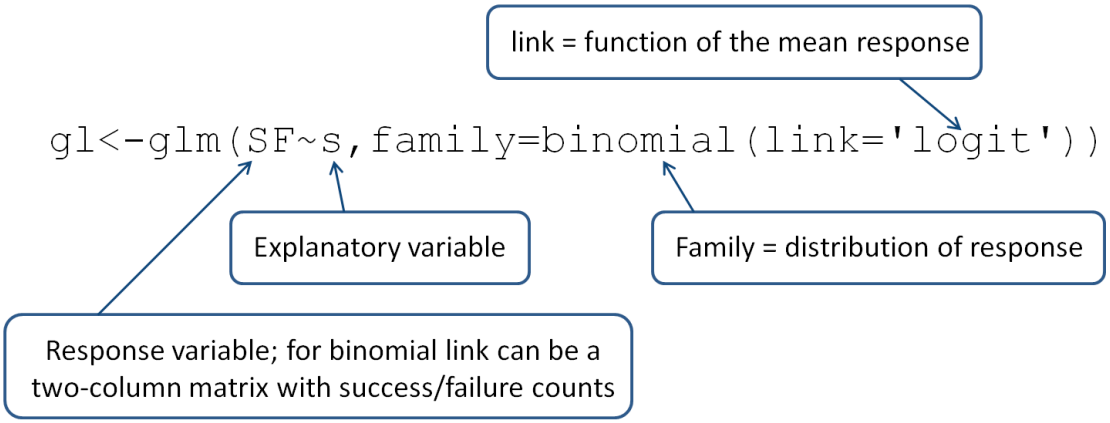
We need to choose the numerical values for snoring levels. For example, for snoring levels `x=[0,2,3,4]` we find, using the MLE approach:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=-3.95+0.52x$$

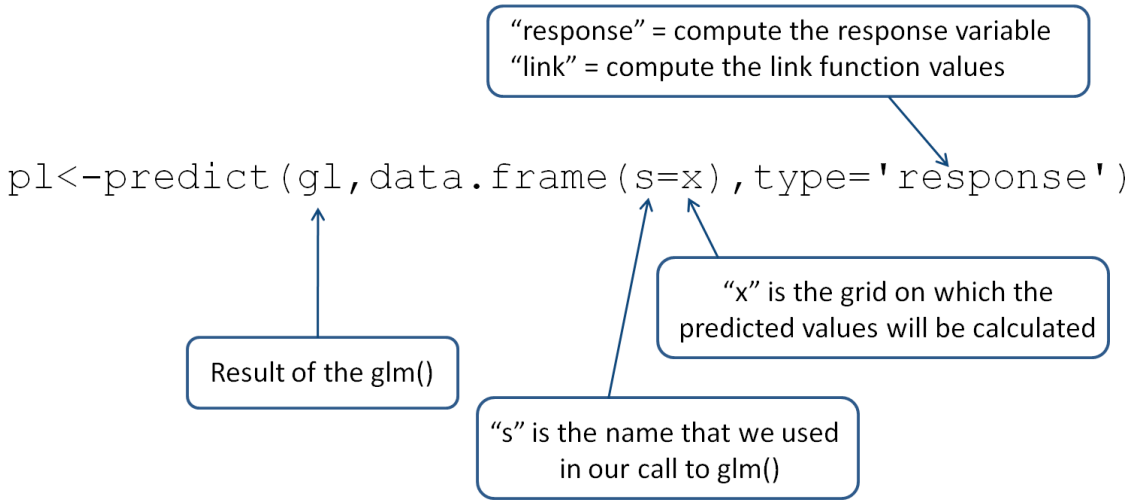
$$probit\left(\pi(x)\right)=-2.09+0.24x$$

The sample code in program `glm_1D.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()`.

Details of **`glm()`:**



Details of **`predict()`:**



Sample R session (file glm.R)

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

# One independent (explanatory) variable
#=======================================

# Set-up
#=======================================
s<-c(0,2,3,4)            # snoring levels
hd<-c(24,35,21,30)       # heart disease counts
n<-c(1355,603,192,224)   # healthy patient counts
SF<-cbind(hd,n)          # matrix of "successes" and "failures"
p<-hd/(n+hd)             # probabilities of "success"

# GLM estimation
#=======================================
gl<-glm(SF~s,family=binomial(link='logit'))
gp<-glm(SF~s,family=binomial(link='probit'))
summary(gl)
summary(gp)

# Response function plot
#=======================================
windows() # new graphical window
plot(c(0,5),c(0,.2),type='n',xlab='Snoring Level',ylab='Proportion')
points(s,p,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) # green line
pp<-predict(gp,data.frame(s=x),type='response')
lines(x,pp,col=4) # blue line
grid()

# Link function plot
#=======================================
windows()
plot(c(0,5),c(-5,-1),type='n',xlab='Snoring Level',ylab='Logit')
points(s,log(p/(1-p)),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)
```